Adapting Pretrained Vision-Language Models in Medical Domains

Liangchen Li

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy in Machine Learning and Machine Intelligence

Clare Hall College

August 2023

I would like to dedicate this thesis to my loving parents.

Declaration

I, Liangchen Li of Clare Hall College, being a candidate for the MPhil in Machine Learning and Machine Intelligence, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose. This dissertation contains 14,998 words, which is fewer than 15,000 words including footnotes, tables, captions and equations.

All experiments were carried out in Python, particularly with PyTorch [PGM+19][1]. The main code base of our ICE pretraining framework is modified from PMC-CLIP [LZZ+23][2], and the training code of BLIP-2 [LLSH23] is modified from LAVIS [LLL+23][3]. For reproduction of MedVInT, we refer to Zhang et al. [ZWZ+23][4]. We use the datasets PMC-OA [LZZ+23][5] and PMC-VQA [ZWZ+23][6] for pretraining, and VQA-RAD [LGBADF18][7] as a downstream task. Pretrained weights of BERT [KT19][8], BiomedCLIP [ZXU+23][9], PMC-CLIP [LZZ+23][5], PubMedBERT [GTC+21][10] and PMC-LLaMa-7B [WZZ+23a][11] are used for model initialisation.

Liangchen Li
August 2023

---

[1] https://pytorch.org/
[2] https://github.com/WeixiongLin/PMC-CLIP
[3] https://github.com/salesforce/LAVIS
[4] https://github.com/xiaoman-zhang/PMC-VQA
[5] https://huggingface.co/datasets/axiong/pmc_oa_beta
[6] https://huggingface.co/datasets/xmcmic/PMC-VQA
[7] https://osf.io/89kps/
[8] https://huggingface.co/bert-base-uncased
[9] https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224
[10] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract
[11] https://huggingface.co/chaoyi-wu/PMC_LLAMA_7B

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Samuel Albanie. His deep insights into cutting-edge research, combined with his rich experience, have been invaluable. I am particularly grateful for his attentive guidance and engaging discussions through the whole research progress. I am also thankful to my co-supervisor, Mingkai Zheng, for the great amount of time and effort he dedicated, and his valuable suggestions that have greatly improved my work. My sincere thanks to both of them for enabling the completion of this research.

Additionally, I wish to extend my special thanks to all my friends at Cambridge/London or in my country, including Guangyu Yang, Yuan Lu, Jiajun He, Zhiqi Zhuang from the MPhil in MLMI program, Weiling Du, Binjie Chen, Caiqi Zhang, Chengzu Li from the MPhil in ACS program, as well as Mandy Chan, Marco Xu, Renjie Zhao, Ling Zhang, for you companionship and mental support through the toughest days, and all the great fun we had. Please forgive me for not being able to list everyone's names here, but know that I'm equally grateful to everyone who has been a part of my journey.

Abstract

With the rise of large-scale vision-language pretraining (VLP), tremendous progress has been achieved in handling complex multi-modal information in general domain. However, large vision-language models specifically tailored for medical domains are far from well-developed due to challenges such as cost, acquisition, privacy concerns, and the intricate nature of medical data. While existing research has delved into constructing better datasets and refining model architectures for medical VLP, there is a noticeable lack of focus on adapting these pretrained models to complex downstream medical tasks. Using medical Visual Question Answering (VQA) as a representative downstream task that is hard for adaptation, this thesis studies the adaptation problem from two points: 1) how can the pretraining target be adjusted to better serve downstream tasks, and 2) how can downstream tasks better align with the pretraining tasks. For the first question, based on our observations that highly similar medical images may harm image feature extraction and fine-grained language supervision is lack for fine-grained discriminative tasks including multiple choice VQA, we propose a Inner Contrastive Enhanced (ICE) Pretraining framework to address such problems, which significantly improves the state-of-the-art on PMC-VQA-test-clean multiple choice questions by 6%. For the second question, we further explore the ICE pretrained model's adaptation to very hard open-ended PMC-VQA questions and a simpler downstream VQA task on VQA-RAD. In the open-ended VQA task, we observe an accuracy-readability trade-off in existing generative models due to poor adaptation, and propose a "first guess answer candidates, then select the best one with ICE" paradigm as well as a simple ensemble trick that significantly improve answer readability while maintaining high accuracy. When adapting to VQA-RAD, we leverage advanced LLMs to restate and augment the closed-ended questions and options to better align with PMC-VQA used for pretraining, and to generate hard textual negatives for open-ended questions. Our approach establishes a new state-of-the-art of 81.82% overall accuracy, implying the great power and flexibility of our ICE pretraining framework and our adaptation methods.

Table of contents

1 Introduction **1**

2 Background **5**

2.1 Vision-Language Pretraining in General Domain 5
2.2 Medical Vision-Language Pretraining Frameworks 6
2.2.1 Contrastive Models 6
2.2.2 Generative Models 9
2.3 Vision-Language Pretraining for Medical VQA 11

3 ICE: Inner Contrastive-Enhanced Pretraining for Medical VQA **14**

3.1 Rethinking Outer Contrastive Pretraining for medical VQA 14
3.1.1 The Pretraining Chain 15
3.1.2 Weaknesses 16
3.2 Large-Scale Inner Contrastive-Enhanced Pretraining 18
3.3 Solving Multiple Choice Medical VQA 21
3.3.1 Network Architecture 21
3.3.2 Experiment Settings 23
3.3.3 Main Results 24
3.3.4 Ablation Study 25

4 Adapting ICE-Pretrained Model to Open-Ended Medical VQA **27**

4.1 The Accuracy-Readability Trade-off in Large Generative Medical VQA Models 27
4.1.1 Understanding Difficulties in Open-Ended Medical VQA 28
4.1.2 The Accuracy-Readability Trade-off 31
4.2 Approach 35
4.2.1 A Reliable GPT-4-Based Evaluation Protocol 37
4.2.2 Improving BLIP-2 Accuracy with ICE-pretrained Model 40
4.2.3 A Simple Ensemble Trick 43

4.3 Experiments 44
4.3.1 Main Results 45
4.3.2 Ablation Study 46

5 Adapting ICE-Pretrained Model to Downstream Medical VQA Task 49
5.1 Dataset and Task Reformulation 49
5.2 Approach 52
5.2.1 Closed-Ended Medical VQA 52
5.2.2 Open-Ended Medical VQA 53
5.3 Experiments 56
5.3.1 Our Settings 56
5.3.2 Evaluation Metric 56
5.3.3 Results and Analysis 57

6 Conclusion, Limitation and Future Work 60

References 62

Chapter 1

Introduction

In recent years, thanks to advancements in hardware computational power, the most prominent trend in the deep learning has been the scaling up of dataset and model sizes. Large language models (LLMs) like GPT-4 with trillions of parameters trained on trillions of tokens [Ope23a] have achieved astonishing capabilities that changed the world, benefiting hundreds of millions of users and significantly boosting social productivity. Meanwhile, large multi-modal models become more and more capable of handling complex information from different modalities including text, images, sound, and video, demonstrating great zero/few-shot ability in downstream tasks including classification, visual question answering (VQA), etc. [RKH+21, JYX+21, WBD+22, LLXH22, LLSH23, ZCS+23, ZZZ+23, YFZ+23, Ope23a, ADL+22, AGG+23, GHZ+23, DLL+23].

However, different from general domain, large vision-language models in medical domains remain underdeveloped due to some notable limitations such as the high cost, difficulty in acquisition, privacy concerns, and fine granularity of medical data. On one hand, even the largest existing medical vision-language dataset PMC-15M [ZXU+23], which has exhausted the largest biomedical database PubMed Central[1], is still orders of magnitude smaller than the large datasets in general domain (e.g., LAION-5B [SBV+22]). On the other hand, even within medical data, there are significant gaps between different subdomains (e.g., CT images and pathological images), so many works only focus on one subdomain as a conservative choice [WZZ+23b, CDW+23, YJT+23, EMDM23, CDH+22, MHWZ23, HSLY21]. These special properties of medical data makes it an extremely challenging problem of how to collect and make best use of the limited available medical data for general medical foundation models.

---

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

While a long line of research has been discussing around how to refine the dataset and model architecture for medical vision-language pretraining and use downstream medical tasks for evaluation, they hardly treat *how to adapt the pretrained models to the downstream tasks* as a serious problem. Chen et al. [CDW+23] classify the downstream tasks into uni-modal (e.g., image classification), cross-modal (e.g., text-image retrieval) and multi-modal (e.g., VQA) tasks and advocate to handle them in a unified framework, as in some following works [ZYY+23, LWZ+23, WZZ+23b]. We argue that although some easy tasks align well with the pretraining task (e.g., cross-modal retrieval aligns well with the image-text contrastive learning), better adaptation strategies for complex tasks including medical VQA are yet to be well-studied, as they are far from the pretraining task. For instance, PubMedCLIP [EMDM23]/BiomedCLIP [ZXU+23] only uses the pretrain visual encoder and has to resort to QCR [ZLF+20]/METER [DXG+22] framework to perform medical VQA, which is clearly not a satisfying way. Among them, only Zhang et al. [ZWZ+23] propose to add an intermediate pretraining step for their large generative model MedVInT on a large medical VQA dataset for better adaptation, which has been shown to be very effective.

In this thesis, using medical VQA as a representative complex downstream task, we further discuss

1. How can the pretraining target be adjusted to better serve downstream tasks?
2. How can we better align the downstream tasks with the pretrained vision-language model for better adaptation?

Note that the two questions start from different perspectives of pretraining/downstream tasks respectively, and both target at better adaptation. For the first question, we propose a new *Inner Contrastive Enhanced (ICE)*-Pretraining framework to train a contrastive model on the large-scale medical VQA dataset PMC-VQA [ZWZ+23] specially designed upon our understandings in the special properties of medical data, which will be demonstrated more powerful than the large generative model in [ZWZ+23] and sets the new state-of-the-art on PMC-VQA multiple choice questions. For the second question, we study adaptation of the ICE-pretrained model to open-ended hard medical VQA and to a simple downstream VQA task. For the former, we leverage the discriminative power of ICE model to improve accuracy of the generated content. For the later, we mainly resort to advanced LLMs [Ope23b, Ope23a] to restate questions and options for better alignment and augmentations. We summarise our main contributions and organisation of this thesis as follows:

- In Chapter 2, we introduce the background of vision-language pretraining, especially existing frameworks in medical domains. We point out that almost all existing contrastive models are pretrained on image-caption pairs that poorly adapt to complex downstream tasks including VQA, and the only generative model pretrained for medical VQA is a suboptimal choice for discriminative (multiple choice) tasks.
- In Chapter 3, we first identify the problem of highly similar medical images in the first stage image-caption pretraining, and then the problem of lack of fine-grained language supervision for the VQA task. We propose a new *Inner Contrastive Enhanced (ICE)*-Pretraining framework for a contrastive model to solve these problems, and improves the state-of-the-art accuracy on PMC-VQA-test-clean multiple choice questions [ZWZ+23] by a large margin of 6% from 42.3% to 48.3%. To our best knowledge, this is the first contrastive model pretrained for large-scale hard medical VQA task.
- In Chapter 4, we study a hard open-ended medical VQA task, PMC-VQA [ZWZ+23]. We first observe an accuracy-readability trade-off and attribute it to the model's limited fitting power and the complexity of the task, and then propose a new pipeline to first predict candidate options and then select the best one using ICE-pretrained model, and also an ensemble trick to achieve good balance in the trade-off. Based on a newly proposed reliable GPT-4-based evaluation protocol, our method significantly increases the answer readability while having good accuracy.
- In Chapter 5, we adapt the ICE-pretrained model to a small downstream VQA task, VQA-RAD [LGBADF18]. For closed-ended questions we use GPT-3.5-turbo [Ope23b] to restate the questions and options in the same format of multiple choice questions in PMC-VQA for better alignment and augmentations. For open-ended questions we generate hard negatives that echoes the fine-grained supervision idea in our ICE framework. Through extensive experiments we show the effectiveness of each proposed technique and it establishes a new state-of-the-art of 81.82% overall accuracy.
- In Chapter 6, we conclude our work and discuss its limitations and future works.

And also refer to Figure 1.1 for the overall organisation.

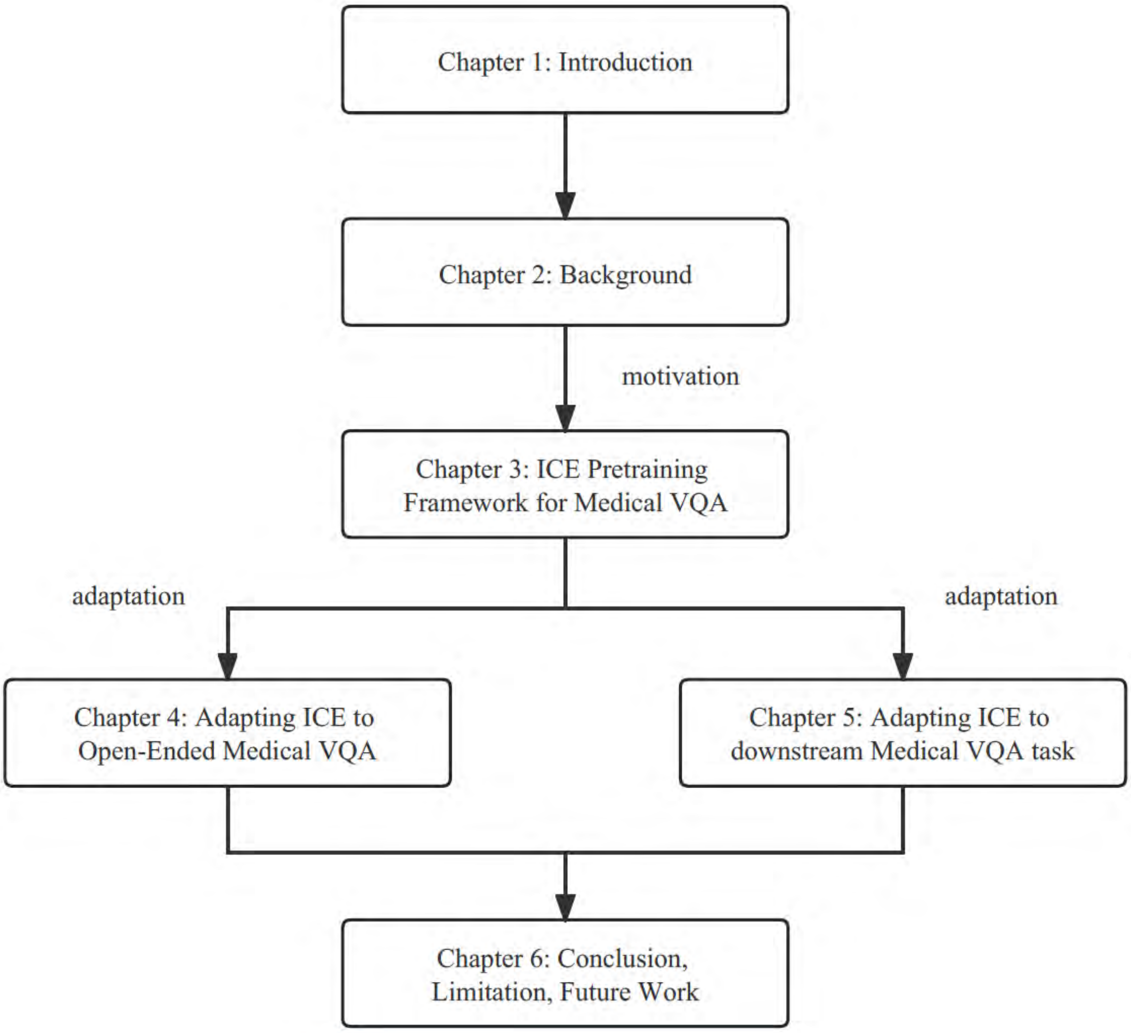

Chapter 1: Introduction
Chapter 2: Background
motivation
Chapter 3: ICE Pretraining Framework for Medical VQA
adaptation
adaptation
Chapter 4: Adapting ICE to Open-Ended Medical VQA
Chapter 5: Adapting ICE to downstream Medical VQA task
Chapter 6: Conclusion, Limitation, Future Work

Fig. 1.1 Overall organisation of this thesis.

Chapter 2

Background

In this chapter, we introduce the background of vision-language pretraining (VLP). Since our main focus is VLP in medical domains, we first give only a brief introduction of VLP in general domain, and defer more details of training targets/architectures adopted in existing VLP frameworks in medical domains to Section 2.2. Finally we narrow it down to pretraining methods for medical VQA, the representative hard-to-adapt downstream task discussed through the thesis.

2.1 Vision-Language Pretraining in General Domain

VLP for representation learning. Compared with supervised [HZRS16] and unsupervised learning [CKNH20, HFW⁺20, GSA⁺20, CTM⁺21], contrastive learning with natural language supervision has been demonstrated to have better generalisation in many downstream tasks [RKH⁺21, LSG⁺21, JYX⁺21, SLT⁺22]. Pretrained on web-scale image-caption pairs [SBV⁺22], the framework uses a simple idea to match the image representations to the representations of their corresponding captions while repelling the others. Many variants have been developed for better performance, such as FILIP [YHH⁺22] that pursuits a patch/token level of image-text alignment and FLIP [LFH⁺23] that randomly masks image patch during training. Beyond learning good image embeddings, more losses including Mask Image/Language Modelling loss [HCX⁺22, KT19] and Image-Text Matching loss are often integrated into the learning object for diverse multi-modal downstream tasks [SHG⁺22, DKG⁺22, LLXH22], e.g., learning a fused representation of image and question for visual question answering (VQA). We defer more details into Section 2.2 under medical VLP background.

Roles of LLMs in VLP. Thanks to the rapid growth of advanced LLMs, vision-language models' text generation ability can greatly benefit from them by pretrained LLMs. For instance, Flamingo [ADL+22, AGG+23] insert learnable attention layers into frozen pretrained language model (Chinchilla-70B [HBM+22]) layers to interact with extracted image information to handle interleaved image-text sequences; BLIP-2 [LLSH23], InstructBLIP [DLL+23] and MiniGPT-4 freeze pretrained LLMs (OPT [ZRG+22], FlanT5 [CHL+22], Vicuna [CLL+23]) and just fine-tune an adaptation layer. Moreover, the most advanced LLMs [Ope23b, Ope23a] are also used for text preprocessing/augmentation. Recent works show high-quality generated textual data are essential for large-scale pretraining [ZCS+23, FKI+23, GZA+23].

Downstream tasks. The pretrained models can be fine-tuned and directly apply to (zero-shot) multiple downstream tasks, such as uni-modal classification, image captioning, cross-modal retrieval (e.g., finding the image that best fits a given description), VQA (usually treated as a classification task over all candidate answers [GKSS+17, CLY+20] or pure open-ended generative task [LSG+21, ADL+22, AGG+23, LLXH22, LLSH23]). We remark that different downstream tasks may focus on different ability of the pretrained model, e.g., while retrieval tasks emphasise contrastive power across modalities, VQA tasks use fused features where the two modalities work in a cooperative manner.

2.2 Medical Vision-Language Pretraining Frameworks

In this section, we revisit existing frameworks that dedicate on large-scale pretraining in medical domains. Particularly, we discuss contrastive and generative models separately,

2.2.1 Contrastive Models

We start from two of the most representative contrastive models pretrained on large-scale medical dataset, BiomedCLIP [ZXU+23] and PMC-CLIP [LZZ+23]. While the former is pretrained on a larger biomedical dataset PMC-15M that contains 15 million image-caption pairs [ZXU+23], the later is pretrained on a smaller (1.6 million image-caption pairs) but more carefully cleaned dataset PMC-OA [LZZ+23]. As shown in Figure 2.1, they both first extract image and text features from a image-caption pair by feeding them into an image encoder and a text encoder. Given the image-caption pairs, pretraining of both the two contrastive models includes the traditional image-text contrastive (ITC) loss as adopted in

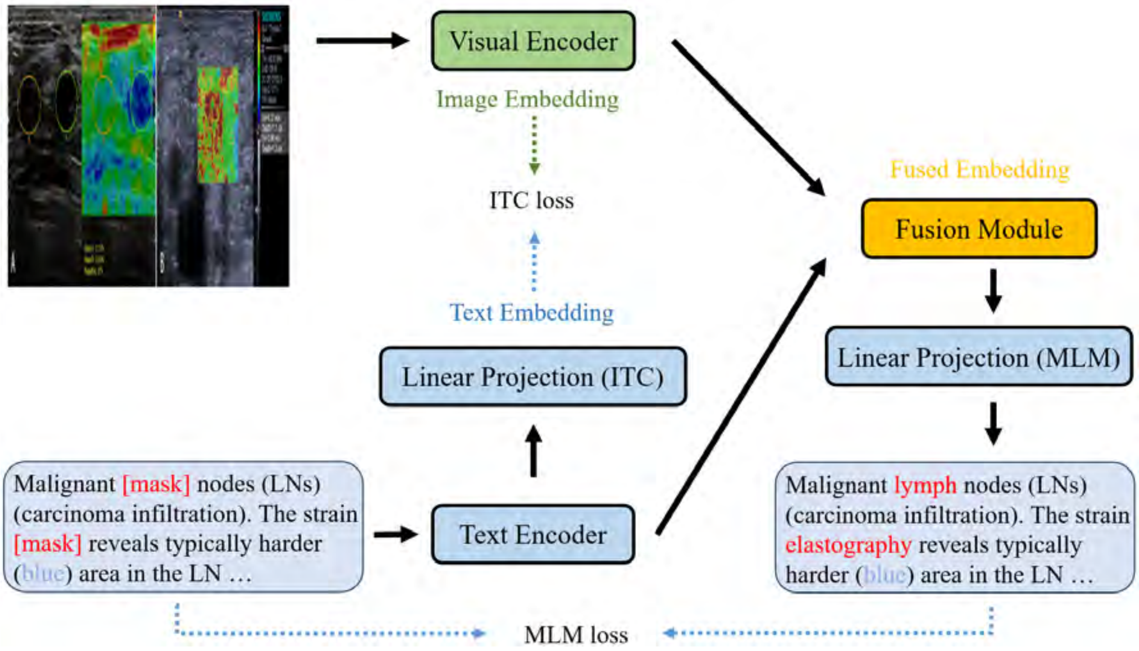

Visual Encoder
Image Embedding
ITC loss
Text Embedding
Linear Projection (ITC)
Fused Embedding
Fusion Module
Linear Projection (MLM)
Malignant [mask] nodes (LNs) (carcinoma infiltration). The strain [mask] reveals typically harder (blue) area in the LN …
Text Encoder
Malignant lymph nodes (LNs) (carcinoma infiltration). The strain elastography reveals typically harder (blue) area in the LN …
MLM loss

Fig. 2.1 Network architecture of PMC-CLIP [LZZ+23]. BiomedCLIP [ZXU+23] shares the same two encoders but without the fusion module and removes the MLM loss.

pretraining general domain CLIP [RKH+21]:

$$\ell_{\mathrm{ITC}} = \frac{1}{2}(\ell_{\mathrm{i2t}} + \ell_{\mathrm{t2i}}) = -\frac{1}{2N}\sum_{i}^{N}\log\frac{\exp(x_i^T y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^T y_j/\sigma)} - \frac{1}{2N}\sum_{i}^{N}\log\frac{\exp(y_i^T x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^T x_j/\sigma)}, \tag{2.1}$$

in which N is the number of pairs contained in a minibatch, x_i/y_i denotes normalised embedding of image/text in the i-th pair in the minibatch extracted by the encoders, and σ is a (learnable) temperature parameter widely adopted in contrastive learning [CKNH20]. Minimising the ITC loss encourages the feature of an image to have higher similarity with the feature of its corresponding caption than other captions appear in the same batch, making it a powerful model for image-text retrieval. Also in a high-level, the captions provide natural language supervision for the images that leads to good image representations if the pretraining dataset is adequately large [RKH+21].

In addition, PMC-CLIP also introduces a Transformer-based fusion module [VSP+17] to fuse the image feature with the text feature, and perform Masked Language Modelling (MLM) loss following BERT [KT19] by predicting the masked tokens in the input caption from the fused embedding as illustrated in Figure 2.1. Specifically, around 15% input caption tokens are replaced with Token [MASK], and then the output of the text encoder, i.e. the input sequence embedding $t \in \mathbb{R}^{\ell \times d}$ (where ℓ is the maximum token sequence length and d is the embedding dimension) is concatenated with a learnable image token embedding $v_0 \in \mathbb{R}^d$ and the image embedding $v \in \mathbb{R}^d$ into $t_{\text{concat}} = [t : v_0 : v] \in \mathbb{R}^{(\ell+2) \times d}$. It is then fed into the fusion model, and the last two rows of the output (corresponding to the position of the image token and the image feature) are removed to get the fused feature, i.e. $t_{\text{fused}} = \Phi_{\text{fusion}}(t_{\text{concat}})[: -2] \in \mathbb{R}^{\ell \times d}$. Finally, we predict the mask tokens over the vocabulary through a linear project linear $p_{\text{MLM}} \in \mathbb{R}^{d \times len(vocab)}$, and punish the Cross Entropy (CE) loss between the predicted tokens and the ground truth:

$$\ell_{\text{MLM}} = \mathbb{E}_{(I,T) \sim D} \left[\mathbf{CE} \left(y^{\text{mask}}, p_{\text{MLM}}(t_{\text{fused}})^{\text{mask}} \right) \right]. \tag{2.2}$$

Through the attention mechanism in the fusion model, minimising this loss enforces the model to learn from both the context and the image to correctly predict the masked tokens, which equips the model with some multi-modal reasoning ability.

Other contrastive models usually include more loss terms (e.g., ITM loss [LLT+22, CDW+23], MIM loss [LLT+22, CDH+22]) into the pretraining objective function and modify the network architecture that better serves the objective function (e.g., introduce a image decoder for MIM loss [CDH+22]), which trade implementation complexity for a marginal performance boost. Besides, they are usually pretrained on a far smaller dataset (e.g., ROCO [PKR+18], which contains only radiology images) that highly constrains generalisation across medical subdomains.

Limitations. Although existing contrastive models have demonstrated strong power in relatively simple uni-modal and cross-modal downstream tasks including uni-modal classification and image-text retrieval, none of them are specially pretrained for hard multi-modal tasks, such as medical VQA that demands strong reasoning ability. For instance, Eslami et al. [EMDM23] and Zhang et al. [ZXU+23] only make use of the pretrained visual encoder and have to resort to previous frameworks QCR [ZLF+20] and METER [DXG+22] to perform medical VQA. Even though a fusion module is introduced in PMC-CLIP so that we can input the questions in the format of "Q: question, A: [MASK]" and expect the model to predict the masked tokens as answers, BERT is unfortunately not good at generative tasks and this fails to take advantage of the superior discriminative ability of a contrastive model (verified in Zhang et al. [ZWZ+23], a pretrained PMC-CLIP shows almost zero ability when doing zero-shot VQA in this way). A fundamental reason behind is the provided pretraining datasets all consist of raw image-caption pairs that mainly model the coarse-grained

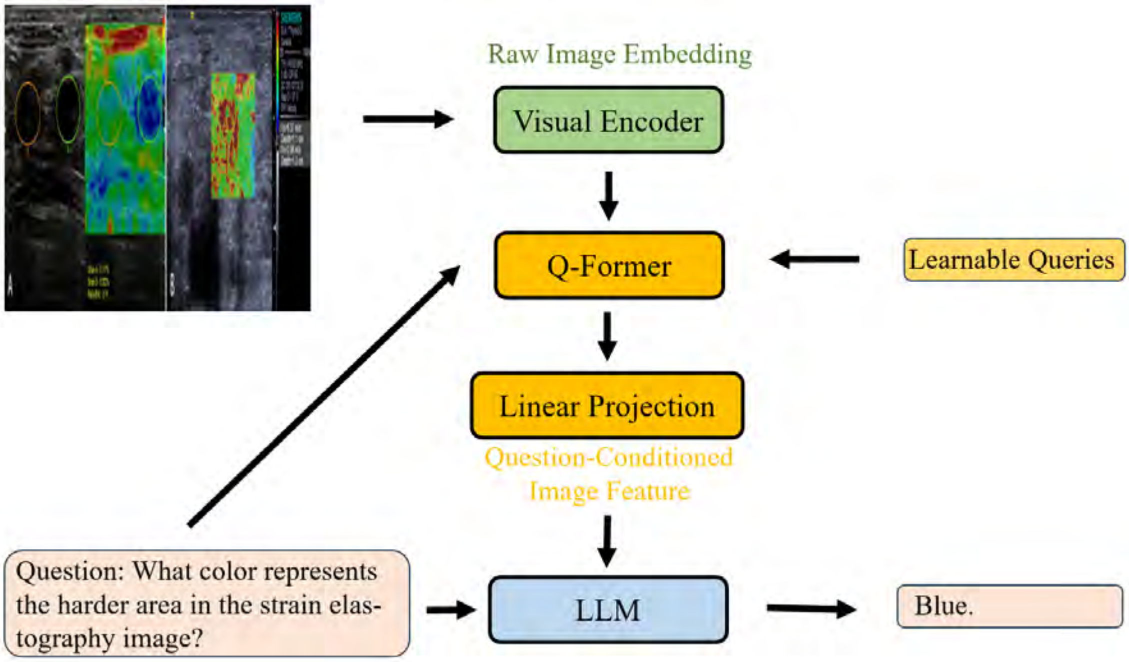

Raw Image Embedding
Visual Encoder
Q-Former
Learnable Queries
Linear Projection
Question-Conditioned Image Feature
Question: What color represents the harder area in the strain elastography image?
LLM
Blue.

Fig. 2.2 Network architecture of BLIP-2 [LLSH23].

matching relationship between the two modalities. However, medical VQA requires much more fine-grained reasoning ability among image-question-answer triplets and discriminative power in multiple choice questions (further explained in Section 3.1), making the existing pretrained contrastive models adapt poorly to such tasks.

2.2.2 Generative Models

Recently, multi-modal generative models in medical domains become significantly more powerful thanks to the advanced improvement of LLMs. Particularly, the most advanced generative models share the BLIP-2 architecture [LLSH23].

As shown in Figure 2.2, the image is input into the visual encoder to extract raw image embeddings. Since Li et al. [LLSH23] argue that not all the raw images features are useful for answering the question, they propose the Q-Former module that takes the image embeddings, 32 learnable queries and the question as input to extract question-conditioned image features. The Q-Former consists of two Transformers [VSP+17], i.e., an image Transformer and a text Transformer, that share the same self-attention layers. Specifically, the queries $q_0 \in \mathbb{R}^{32\times d}$ is concatenated with embeddings of the question $q \in \mathbb{R}^{\ell\times d}$ where ℓ is the question length and d is the embedding dimension, and the $32+\ell$ long sequence is input into the image Transformer. In order to interact with the raw image embeddings, a cross-attention layer followed by a feed forward layer is added after the self-attention layer in every Transformer block, and the last hidden state of the final output is used as the question-conditioned image feature. Note that in the first stage of BLIP-2 pretraining, only the learnable queries are input into the image Transformer and captions are input into the text Transformer in the Q-Former, and it is trained with ITC, ITM and LM loss. Since we focus on either its applications in VQA tasks (fine-tuning on a pretrained BLIP-2 model) or models that merely adopt the network architecture, we refer readers to Li et al. [LLSH23] for the pretraining details.

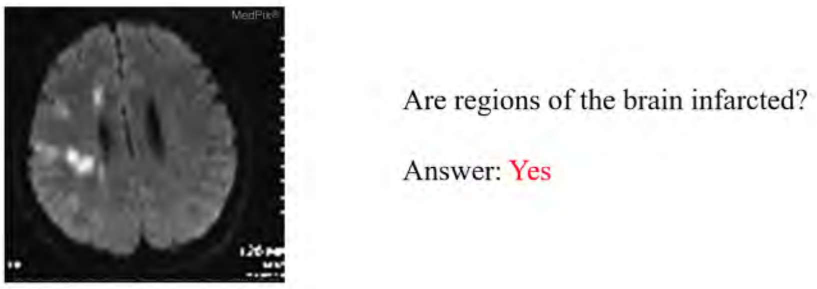
Having been extracted from Q-Former, the image feature goes through a linear projection layer to align with a pretrained LLM. The projected image embedding is then concatenated with the question token embeddings as input to the LLM to predict the output in an autoregressive way. Standard teacher-forcing method [WZ89] is used to compute the language modeling (LM) loss:

$$\ell_{LM} = -\sum_{\ell=1}^{L} \log P(w_t | w_1, w_2, \cdots, w_{t-1}), \tag{2.3}$$

where $P(w_t | w_1, w_2, \cdots, w_{t-1})$ denotes the probability of the next token w_t predicted by the LLM given previous tokens $w_1, w_2, \cdots, w_{t-1}$.

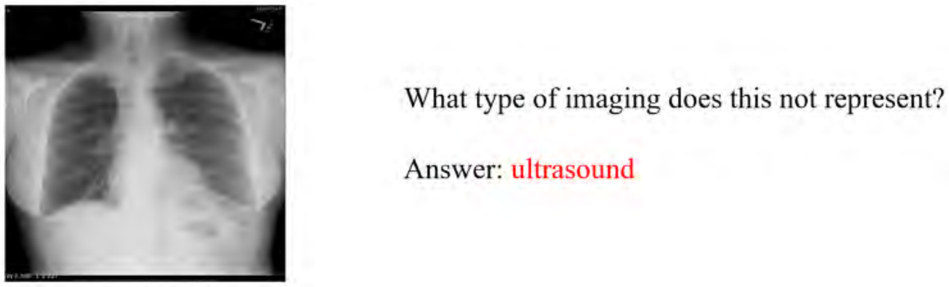
Among existing SOTA multi-modal LLMs in medical domains including BiomedGPT [ZYY+23] and LLaVA-Med [LWZ+23] that are pretrained on large-scale datasets across many medical subdomains or even different medical tasks, only the family of MedVInT models [ZWZ+23] are specially pretrained on PMC-VQA [ZWZ+23], the largest medical VQA dataset. The dataset comprises 227k medical image-question pairs, and instead of freezing the LLM weights as in Li et al. [LLSH23], Zhang et al. use LoRA [HWAZ+22] to fine-tune the LLM in BLIP-2 framework for better fitting ability in pretraining. As a result, the MedVInT family achieves state-of-the-art performance in all of the evaluated downstream medical VQA benchmarks.

Limitations. Despite the success in general domain VQA tasks achieved by multi-modal LLMs such as BLIP-2 as well as extremely large dataset such as LAION-5B [SBV+22], medical VQA tasks are much more challenging, as the questions and medical images can be very fine-grained and large-scale medical VQA datasets of high-quality is extremely difficult and expensive to create, which makes generalisation and even fitting the dataset a hard problem. As reported in Zhang et al. [ZWZ+23], the PMC-VQA dataset itself remains a very challenging benchmark for MedVInT models, even when LoRA is adopted


Are regions of the brain infarcted?

Answer: Yes

(a) closed-ended


What type of imaging does this not represent?

Answer: ultrasound

(b) open-ended

Fig. 2.3 Examples of closed-ended and open-ended medical VQA questions. From VQA-RAD [LGBADF18].

to release part of their fitting ability. Consequently, the highly demanded generative power of free-form text in real applications is at the cost of controllability over the generated content. As will be demonstrated in Section 4.1, existing generative models still struggle in hard open-ended medical VQA tasks, and for multiple choice questions the generated answers may be none of the provided options or even human-unreadable. Moreover, LLMs integrated in the generative models makes the pretraining, fine-tuning and inferring significantly computationally expensive and time-consuming. These reasons together make them no better an option than contrastive models, especially for multiple choice medical VQA questions. Therefore, we mainly discuss pretraining contrastive models in Chapter 3, and more discussions on generative models will be left to Chapter 4 where they are necessary for open-ended problems.

2.3 Vision-Language Pretraining for Medical VQA

Medical VQA questions are either closed-ended or open-ended [LGBADF18]. As the examples shown in Figure 2.3, the answer to a closed-ended question is within a known, limited range (e.g., in Figure 2.3a the answer should be either "yes" or "no"), while the answer to an open-ended question can be any free-form text. However, this is often treated as a classification task over all possible candidates within a VQA dataset, and we defer more discussions on this formulation to Section 5.1. In retrospect of previous methods that aim to solve the medical VQA task with pretraining, the long history can be roughly classified into three stages.

The MEVF [NDN$^+$19] and CR [ZLF$^+$20] framework are probably the earliest two frameworks that adopt pretraining. In this first stage, pretraining are conducted within one modality, i.e., the text encoders and image encoders are pretrained separately. Specifically, they pretrain two visual encoders using MAML [FAL17] by viewing VQA-RAD as a image classification task and CDAE [MMCS11] on manually collected 12k unlabeled medical images, and use GloVe [PSM14] to extract question embeddings. They use BAN [KJZ18] and SAN [YHG$^+$16] as the fusion network to fuse all the features into a fused feature, and then it is fed into a classifier and trained with Equation 5.1. Later, the CR framework is built on this, but they introduce a bifurcation in the model to treat closed-ended questions and open-ended questions differently, considering the different reasoning difficulty in answering them (using a more complex reasoning module for open-ended questions). These classic frameworks are widely used in later works including PubMedCLIP [EMDM23], where only the visual encoder is substituted with a newly pretrained one.

In the era following the emergence of CLIP [RKH$^+$21], during which natural language supervision becomes the new popularity and relatively large-scale medical multi-modal datasets emerge [PKR$^+$18, SWM$^+$20, JPG$^+$19], the second stage mainly focuses on improving the way of fusing features from the two modalities in terms of both training objective and fusion module in VLP. On the loss function side, MMBERT [KBM$^+$21] uses MLM loss to train a Transformer for fusion; based on the understanding in fine-granularity of medical tasks, GLoRIA [HSLY21] learns to match both global and local image/text features; inspired from MAE [HCX$^+$22], M3AE [CDH$^+$22] and M2I2 [LLT$^+$22] include MIM loss for fine-grained image reconstruction; M2I2, RAMM [YJT$^+$23] and PTUnifier [CDW$^+$23] further introduce ITM loss into the training target for better fused embeddings. On the fusion module side, BiomedCLIP [ZXU$^+$23] adopts METER [DXG$^+$22] as a better fusion module than the CR framework; PTUnifier introduces pseudo visual/textual token pools for a uniformed training framework of uni/cross/multi-modal tasks; in RAMM [YJT$^+$23], Yuan et al. specially design a retrieval fusion module to fuse auxiliary image-text pairs with the original ones in their retrieval augmentation strategy. Note that these methods mainly aim to get good fused representations of the two modalities, and the fused embeddings are still sent into a classifier in downstream VQA tasks.

The third stage emphasises more on constructing large-scale multi-modal medical datasets. Most of the works above in the second stage, also including CR+CP [LZXW22] and PubMedCLIP [EMDM23], choose to pretrain on datasets such as ROCO [PKR+18] and MIMIC-CXR [JPG+19] that mainly consist of radiology images. However, the scale and lack of diversity are clearly not satisfying, driving researchers to build larger and more diverse datasets. Subsequently, RAMM [YJT+23], PMC-CLIP [LZZ+23] and BiomedCLIP [ZXU+23] build PMCPM, PMC-OA and PMC-15M that contains 0.4M/1.6M/15M image-caption pairs respectively from the Open Access subset of PubMed Central[1]. Most recently, Zhang et al. [ZWZ+23] construct a large-scale medical VQA dataset PMC-VQA by prompting ChatGPT [Ope23b] to raise questions and answers based upon captions. It is worth mentioning that although some previous works also advocate to treat medical VQA as a pure generative task instead of classification [RZ20, SPR21, VSDN+23] (discussions on which are skipped due to poor performance/insignificant contributions, e.g., simply fine-tuning existing pretrained models on small medical VQA datasets using existing fine-tuning techniques), PMC-VQA is the first medical VQA pretraining dataset large enough to support good performance under the generative formulation [ZWZ+23].

---

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Chapter 3

ICE: Inner Contrastive-Enhanced Pretraining for Medical VQA

The limitations of existing medical VLP frameworks discussed in Section 2.2 motivate us to explore *contrastive models pretrained on/for large-scale hard medical VQA tasks*. One of our key understandings is that, *pretraining strategies can be tailored for different downstream tasks for better performance*, and it is important to rethink what kind of ability does the downstream tasks truly want to benefit from the pretraining. In this chapter, we first point out that for *some* downstream medical tasks including VQA, the traditional contrastive pretraining, which we call *outer contrastive pretraining*, may expose some weaknesses. We then introduce our ***I**nner **C**ontrastive-**E**nhanced (ICE) pretraining* strategy, which is specially designed to overcome the weaknesses for such downstream tasks. We further demonstrate such pretraining scheme is particularly powerful in hard medical VQA tasks.

3.1 Rethinking Outer Contrastive Pretraining for medical VQA

As pointed out before, existing contrastive models are all pretrained on image-caption pairs and directly adapt to downstream VQA tasks, and it is important to introduce pretraining on PMC-VQA as an intermediate stage for better adaptation [ZWZ$^+$23]. In this section, we first analyse what abilities do we need from each stage of the pretraining chain, and then rethink the potential weaknesses of traditional contrastive learning framework in offering such abilities.

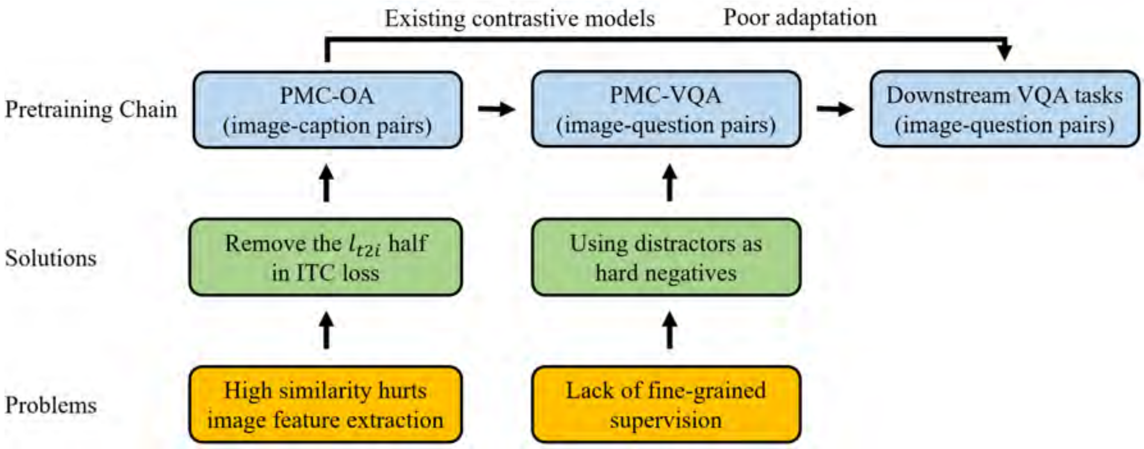

Existing contrastive models
Poor adaptation
Pretraining Chain
PMC-OA (image-caption pairs)
PMC-VQA (image-question pairs)
Downstream VQA tasks (image-question pairs)
Solutions
Remove the l_{t2i} half in ITC loss
Using distractors as hard negatives
Problems
High similarity hurts image feature extraction
Lack of fine-grained supervision

Fig. 3.1 Summary of our idea. In the first pretraining stage on image-caption pairs, we remove the ℓ_{t2i} half in ITC loss to avoid contrasting highly-similar images that can hurt the representation learning. In the second stage on image-question pairs, we use the distractors as hard negatives to increase the model's fine-grained discriminative power.

3.1.1 The Pretraining Chain

Summarised in Figure 3.1, the overall pretraining chain is: **PMC-OA (image-caption pairs) → PMC-VQA (image-question pairs) → downstream medical VQA tasks (image-question pairs)**. We zoom into the two pretraining stages for a detailed discussion on how they serve the downstream task:

1. **Pretraining on image-caption pairs**. In this stage, the image-text contrastive learning in Equation 2.1 plays the most important role. We emphasise that unlike text-to-image retrieval, multi-modal tasks including VQA focus on the fusion of features from the two modalities and have no special requirement on the discriminative power among images given a text (e.g., contrasting images given a question/answer makes no sense and thereby not needed in VQA tasks). Therefore, this stage mainly aims to learn good representations of images (and texts) for fusion, which shares the same understanding in many previous works including PubMedCLIP [EMDM23], BiomedCLIP [ZXU+23] and MedVInT [ZWZ+23] that only detach the pretrained visual encoder in this stage for image embeddings.

2. **Pretraining on image-question pairs**. In this stage, the pretraining task aligns with the downstream tasks, thus providing every ability needed. Note that the usual way to do VQA is to contrast *all candidate answers* against the correct answer, as will be introduced in Section 5.1.

3.1.2 Weaknesses

Now we rethink whether the two pretraining stages best provide the wanted ability, particularly in medical domains. Unlike in general domains, we notice that medical tasks are often highly professional and fine-grained: while everyone can tell "whether the man or woman is wearing glasses in a photo"[1], only medical experts can diagnose a certain kind of disease from tiny clues in radiographs or lesion sections. Given these special properties in medical domains, the situation may change a little when directly applying the traditional ITC loss in Equation 2.1 to pretraining in medical domains. We also call it *outer contrastive pretraining* as it contrasts an image-caption pair against all other negative pairs in a minibatch. Here we zoom into two concrete properties of vision-language pretraining in medical domains that may expose the weaknesses of outer contrastive pretraining.

Stage I: High image similarity hurts image feature extraction. The appearance, size, color, and location of objects can vary greatly in general natural images, so 1) very similar images hardly appear, and 2) similar images often have consistent descriptions. However, medical images often reflect subtle anatomical structure differences or pathological changes, therefore 1) highly similar images within some specific subdomains widely exist, and 2) their corresponding captions can be very different.

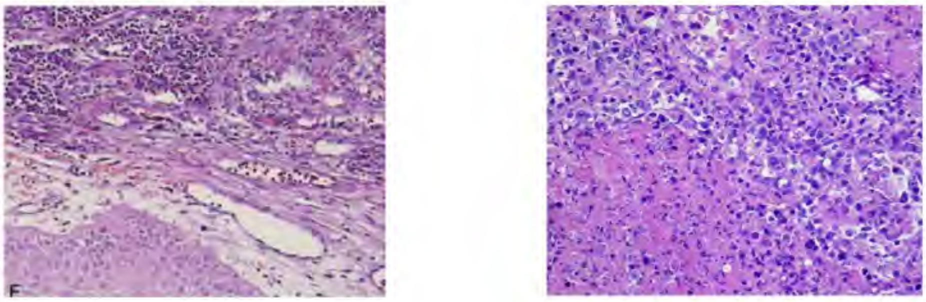
To illustrate such property and difficulties it brings, as an example, we manually pick out some histopathological images using H & E staining from PMC-OA [ZWZ$^+$23], and feed them into the state-of-the-art contrastive model BiomedCLIP [ZXU$^+$23][2], pretrained on the largest PMC-15M dataset with ITC loss only, for a simple retrieval test. In the first example (Figure 3.2a), an intracocular malignant tumor and a kind of lymphoma are wrongly identified as each other with high confidence, and in the second example (Figure 3.2b) a kind of pneumonia is wrongly identified as a kind of cancer. It turns out that even the most powerful contrastive model struggles to identify either the tissue or disease from a single histopathological image due to their high similarity, so do human experts.

Recall that our main purpose in the first pretraining stage is to get good image embeddings for fusion. Previous works in general domain contrastive learning have shown that dissimilar positive examples created with strong augmentations can hurt the representation learning [ZYW$^+$21], while here in the same spirit, widely existing highly similar negatives such as

---

[1] Example quoted from Goyal et al. [GKSS$^+$17]. See also https://visualqa.org/.

[2] We use its online demo at https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224.



PMC4616708_F2_435835

Left caption: 4.9% √
Right caption: 95.1% ×

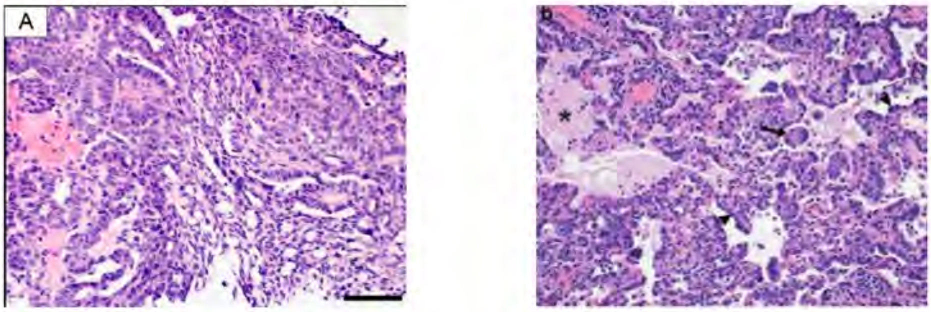
PMC7506053_F3_258416

Left caption: 94.7% ×
Right caption: 5.3% √

Left: Histopathological examination (H&E stain 200×) of the fluid taken from intraocular surgery demonstrates tumor-like cells, suggesting an intraocular malignant tumor.

Right: Dense infiltration of the brain composed of Reed Sternberg cells small lymphocytes and macrophages corresponding to the classical Hodgkin lymphoma of the central nervous system mixed cellularity (MCcHL) and positive immunostaining…

(a) example 1



PMC6547356_ijms-20-02131-f001_488723

Left caption: 100.0% √
Right caption: 0.0% ×

PMC7241103_F4_175648

Left caption: 99.2% ×
Right caption: 0.8% √

Left: Photomicrographs of selected cases showing high-grade serous carcinoma. (A) Patient #17 high-power magnification of pre-NACT omental biopsy showing solid clusters of malignant cells with high purity and no intervening stroma.

Right: Histological changes in lungs of rhesus macaques on 7 dpi. Focal interstitial pneumonia in lungs of a control animal. The area in the black box is magnified in panel b. Interstitial pneumonia with edema (asterisk) type II pneumocyte hyperplasia…

(b) example 2

Fig. 3.2 Example image-pairs of highly similar images picked from PMC-OA [LZZ+23]. BiomedCLIP [ZXU+23] struggles in discriminating them by retrieval, which may also bring confusion in outer contrastive learning.

the histopathological images shown above can be too difficult for the model to learn good image representations.

Stage II: Insufficiently fine granularity of language supervision. Previous works in general domain VLP have demonstrated the importance of fine-grained alignment between images and captions [ZCS+23]. As for contrastive models in medical domains, fine-grained discriminative power is even more demanded as the task (e.g. questions in medical VQA) can be very specific.

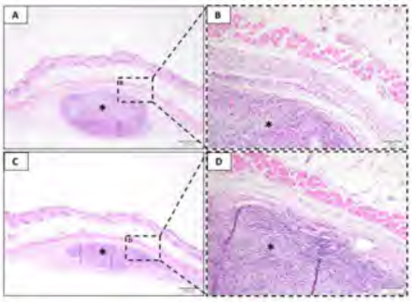
To examine whether existing contrastive models owns such ability, we randomly select some image-(multiple choice) question pairs from PMC-VQA [ZWZ+23], and we restate each of the options into a declarative sentence and again feed them into BiomedCLIP [ZXU+23] for retrieval evaluation. As demonstrated by the examples in Figure 3.3, even the most powerful contrastive model shows nearly zero ability in fine-grained discriminative tasks. In fact this is totally understandable, as there is completely no guarantee of explicit fine-grained contrastive information reliably provided for such problems. For instance, to correctly answer the question in Figure 3.3b, we may also need the sonographic image showing hyperchoic area of the same patient and corresponding caption to tell the model their differences in outer contrastive pretraining, which is very unlikely to happen. In other words, we lack hard textual negatives in both the two pretraining stages for fine-grained discriminative ability.

3.2 Large-Scale Inner Contrastive-Enhanced Pretraining

The examples we discussed above suggest: 1) one should be careful when contrasting similar images to different captions in the first pretraining stage; 2) fine-grained hard textual negatives are highly demanded for fine-grained discriminative tasks including medical VQA, but are missing in the pretraining chain. To this end, for each image-caption pair (i,t), we propose to manually modify the caption t into M false distractors $t_1', t_2', \cdots, t_M'$, and contrastive the correct pair (i,t) against all other negative pairs $(i,t_1'), (i,t_2'), \cdots, (i,t_n')$. Based on these pairs, we introduce *inner contrastive loss* to do such contrastive learning:

$$\ell_{\textbf{inner}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(x_i^T y_i/\sigma)}{\exp(x_i^T y_i/\sigma) + \sum_{m=1}^{M} \exp(x_i^T {y_{im}}'/\sigma)}, \tag{3.1}$$

in which x_i/y_i denotes the features of the i-th original image-caption pair, and ${y_{im}}'$ denotes the feature of the m-th false distractor. By minimising the inner contrastive loss, we encourage the features of (i,t) to be more and more similar while encourage the features of other negative



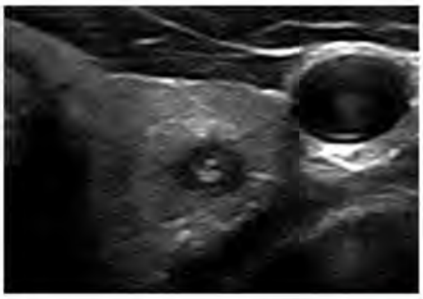
PMC7765348_membranes-10-00423-f004

What is indicated by an asterisk (*) in the photomicrographs?

A: 0.7% √
B: 11.4% ×
C: 29.3% ×
D: 58.5% ×

A. The area occupied by the membrane is indicated by an asterisk (*) in the photomicrographs.
B. The area occupied by the cells is indicated by an asterisk (*) in the photomicrographs.
C. The area occupied by the blood vessels is indicated by an asterisk (*) in the photomicrographs.
D. The area occupied by the extracellular matrix is indicated by an asterisk (*) in the photomicrographs.

(a) example 1



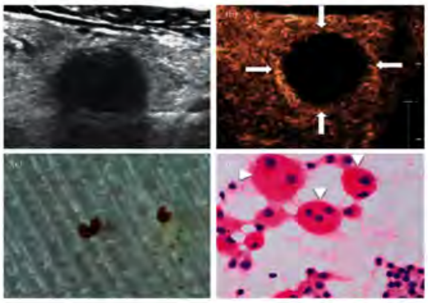
PMC7227110_Fig2

What did the sonographic image of MTMC show?

A: 64.8% ×
B: 20.3% √
C: 9.2% ×
D: 5.7% ×

A. The sonographic image of MTMC shows hyperchoic area.
B. The sonographic image of MTMC shows hypoechoic area.
C. The sonographic image of MTMC shows hyperintense area.
D. The sonographic image of MTMC shows hypointense area.

(b) example 2



PMC9068332_fig3

What is the appearance of the nodule in CEUS?

A: 19.6% ×
B: 71.2% ×
C: 4.4% √
D: 4.8% ×

A. The nodule exhibits regular annular enhancement in CEUS.
B. The nodule exhibits irregular heterogeneous enhancement in CEUS.
C. The nodule exhibits no enhancement in CEUS.
D. The nodule exhibits homogeneous enhancement in CEUS.

(c) example 3

Fig. 3.3 Example image-questions randomly selected from PMC-VQA [LZZ+23] and options are restated into declarative sentences. BiomedCLIP [ZXU+23] shows nearly zero ability in fine-grained discriminative tasks.

pairs to be more and more dissimilar. Since this method does not include interaction with outer examples as negative pairs, we name it ***I**nner **C**ontrastive-**E**nhanced (ICE) pretraining*.

This learning framework enjoys two major advantages:

1. Since each image-caption pair is used as a singleton, it naturally avoids the potential risk of contrasting two similar images in outer contrastive pretraining that has been discussed in Section 3.1.2. This is particularly suitable for those multi-modal downstream tasks that lay more emphasis on fusion of information from the two modalities where the discriminative power among images given a text is not necessary, including medical VQA.

2. The false distractors can provide explicit fine-grained language supervision. Since we do not specify the format of the false textual distractors, they can definitely be hard negatives that only one or few keywords are intentionally falsified or adjusted according to the requirements of downstream tasks, and can also be easily created, e.g., using advanced LLMs like GPT-4 [Ope23a]. The extractors in each of the examples in Figure 3.3 set a good example of how they can be created.

Remark 1 (degenerated inner contrastive loss). If the second half of the outer contrastive loss is removed and the false distractors $t'_1, t'_2, \cdots, t'_M$ are assigned to captions in other (negative) image-caption pairs in a minibatch, then both losses degenerate to the same (which we call *degenerated inner contrastive loss*, the first half in Equation 2.1).

Remark 2 (inner contrastive learning in the pretraining chain). There seems to be contradiction: while we condemn highly similar images for hurting the representation learning, we also advocate to use hard negatives on the text side. However, we emphasise that these are for different pretraining stages. In the first stage where getting a relatively good visual (and text) encoder is our main target, we will show simply using the degenerated inner contrastive loss (without any hard textual negatives) can boost performance in Section 3.3.4, and will also show directly adding fine-grained language supervision as hard negatives when training from scratch can also hurt the representation learning in Section 5.3.3. Therefore, we defer the solution to the fine-grained supervision problem to the second pretraining by using the false distractors in multiple choice questions are hard textual negatives, where the fine-grained supervision then become critical but non-harmful based on a good initialisation, as will be verified in Section 3.3.4 and 5.3.3. Our overall idea is summarised in Figure 3.1,

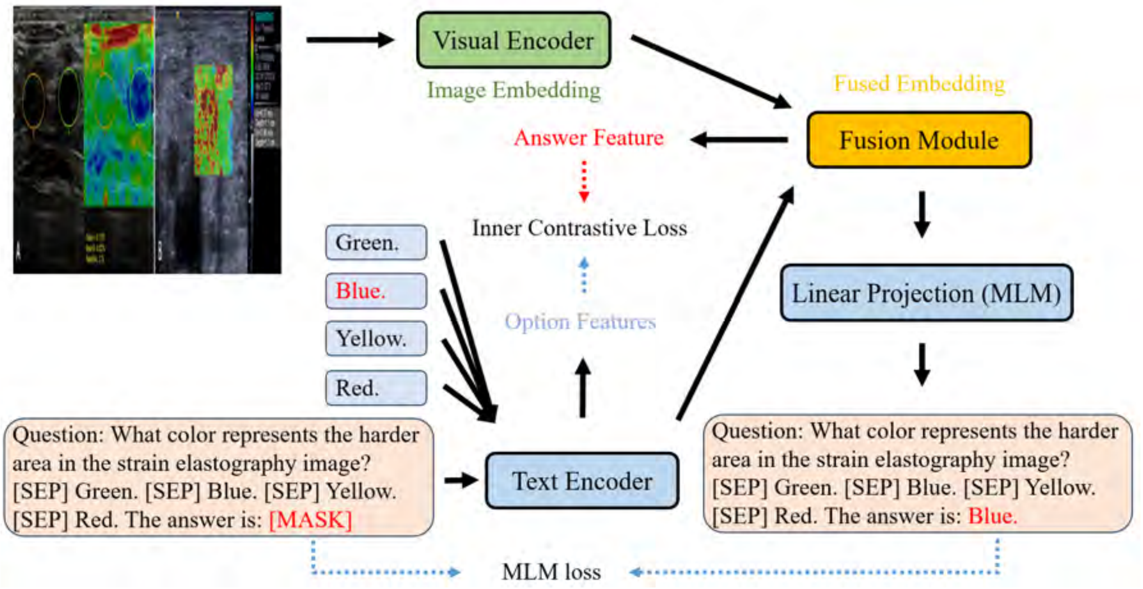

Visual Encoder
Image Embedding
Fused Embedding
Fusion Module
Answer Feature
Inner Contrastive Loss
Option Features
Linear Projection (MLM)
Green.
Blue.
Yellow.
Red.
Question: What color represents the harder area in the strain elastography image? [SEP] Green. [SEP] Blue. [SEP] Yellow. [SEP] Red. The answer is: [MASK]
Text Encoder
Question: What color represents the harder area in the strain elastography image? [SEP] Green. [SEP] Blue. [SEP] Yellow. [SEP] Red. The answer is: Blue.
MLM loss

Fig. 3.4 Network architecture used in ICE pretraining for medical VQA.

which clearly illustrates the main problems and solutions in each pretraining stage of the chain.

3.3 Solving Multiple Choice Medical VQA

In this section, we demonstrate how ICE pretraining can boost model performance in multiple choice medical VQA as a representative multi-modal task in medical domains. We first introduce the network architecture for this task and then report and analyse the results.

3.3.1 Network Architecture

Since the inner contrastive loss shares a similar form with the outer contrastive loss, one can simply adopt the classic contrastive model architecture in Figure 2.1 for ICE pretraining. For multiple choice medical VQA, since the input is no longer image-caption pairs but image-question-option-answer quadruplets, some modifications to the network architecture are needed.

Figure 3.4 demonstrates the overall modified architecture. First, the question is concatenated with the list of options, and further concatenated with the prompt "The answer is: " and finally a sequence of [MASK] tokens, and then fed into the text encoder to extract question

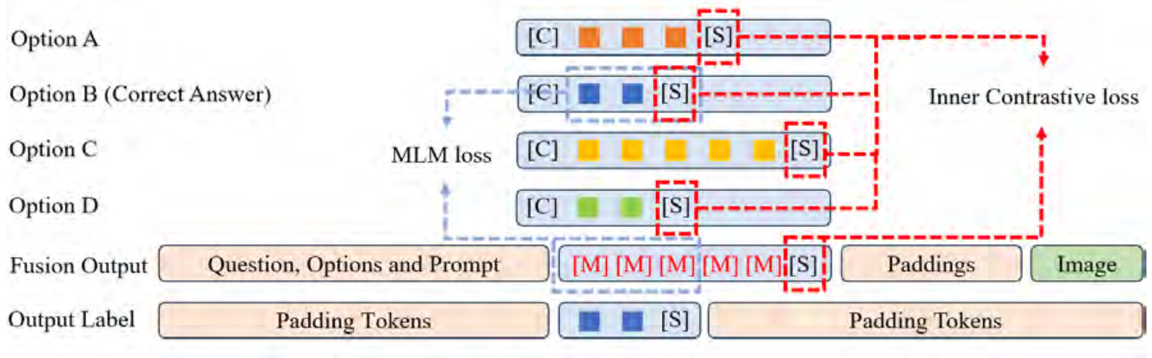


Fig. 3.5 Input and output of the network, in which [S], [M], [C] denotes [SEP], [MASK] and [CLS] tokens respectively.

embeddings. [SEP] tokens are inserted between the question and each option to separate them. Since we do not know how many [MASK] tokens does the ground truth answer take, we use as many [MASK] tokens as the option with the most tokens out of all the options to ensure the correct answer can fit into the blank, as illustrated in Figure 3.5. Same as introduced in Section 2.2.1, the question embeddings is concatenated with image embeddings to be fed into the fusion model to get fused embeddings. Simultaneously, all options are also fed into the text encoder to compute option features, which is set to be the feature at the position of the [SEP] token in the encoder output. Since medical VQA is a multi-modal task, we also use the feature at the position of the last [SEP] token to represent the answer feature to match with the option features instead of directly using the image feature for ICE pretraining. Also note that MLM loss is only applied to the length of the correct answer instead of the whole mask region, as illustrated in the blue dashed squares in Figure 3.5.

While in the original contrastive model architecture the output of the text encoder goes through a linear projection to better align the text feature with the image feature, we simply remove it when extracting option features. One reason is that both the option features and the answer feature lie in the text modality, so no particular effort need to be made for alignment within the one modality. We also clarify that we only do not use the linear projection layer to extract option features, which means it can be re-added to the architecture in Figure 3.4 when needed, e.g., when we still want to use ITC loss together with the inner contrastive loss.

Remark. In this design, x_i in Equation 3.1 represents the answer feature, and y_i and ${y_{im}}'$ represent the features of the correct answer and the other $m = 3$ false options. Since the false distractors in multiple choice medical questions are often semantically closely related to the correct answer, they naturally severe as wanted hard negatives that provide fine-grained supervision.

3.3.2 Experiment Settings

Dataset. We use PMC-VQA [ZWZ+23] as our pretraining dataset in the second stage, which is currently the largest VQA dataset in medical domains containing 227k image-question pairs. We follow the original dataset split. Particularly, Zhang et al. [ZWZ+23] hold a clean subset of 2k test examples whose quality has been manually verified to be good. This dataset is created by prompting ChatGPT [Ope23b] to ask questions about the captions PMC-OA [LZZ+23] and provide the correct answer and 3 other distractors for each question. Although complex filters have been trained to further clean the dataset, e.g., removing questions that do not rely on the image/have to resort to additional information beyond the image to answer, we remark it is still super noisy. This with its great diversity and complexity of both images and questions (see Figure 3.3 for a taste) together make it a good large-scale pretraining dataset but also an extremely challenging benchmark, as concluded by Zhang et al. Since in the second stage our model initialises from PMC-CLIP weights pretrained also on PMC-OA [LZZ+23], we perform a deduplication in PMC-OA to make sure that none of the original images or subfigures of the images in PMC-VQA test samples appear in PMC-OA and re-pretrain PMC-CLIP on the deduplicated dataset.

Training. We choose ViT-Base [DBK+21] as visual encoder and resize the images to 224×224 for input, and use a 4-layer Transformer as the fusion module. For PMC-CLIP pretraining, we follow the settings in Lin et al. [LZZ+23] to train the visual encoder and fusion model from scratch but initialise the text encoder from PubMedBERT [GTC+21], and *use the degenerated inner contrastive loss (de-inner for short)* as in our design. For ICE pretraining on PMC-VQA, we adopt the proposed inner contrastive loss, and also keep the MLM loss in typical contrastive learning as well as de-inner loss between the image and question pairs. Although we admit it make little sense to contrast questions given an image, we will show in the following ablation study that, surprisingly, it may still help improve model performance. Therefore, our total training objective becomes

$$\ell_{\text{ICE}} = \lambda_1 \ell_{\text{inner}} + \lambda_2 \ell_{\text{MLM}} + \lambda_3 \ell_{\text{de-inner}}. \tag{3.2}$$

Following Lin et al. [LZZ+23], we keep equal weights of the MLM and de-inner loss, i.e., setting $\lambda_2 = \lambda_3 = 0.5$, and assign much higher weight $\lambda_1 = 2.0$ to our inner contrastive loss to enhance fine-grained discriminative power. As mentioned above, our ICE model initialises from the pretrained PMC-CLIP weights, and is further trained on PMC-VQA for 10 epochs. We use AdamW optimiser [LH19] and adopt cosine annealing learning rate schedule [LH17] with maximum learning rate 1×10^{-4} and 500 warm up steps. We use a batch size of 128 and maximum text length 128.

Baseline methods. Beside zero-shot performance evaluated on BLIP-2 [LLSH23] and OpenFlamingo [AGG+23] pretrained on large-scale natural image-caption datasets, the only baseline method we are able to compare here is the MedVInT model family [ZWZ+23] since PMC-VQA is released very recently, and we leave comparison with other medical VQA models to an easier downstream medical VQA task in Section 5.3.3. As introduced in Section 2.2.2, MedVInT concatenate Q-Former with a LLM to boost the model's text generation ability. Similar to us, instead of training from scratch, they try pretrained language backbones including PubMedBERT [GTC+21], LLaMa [TLI+23] and PMC-LLaMa [WZZ+23a] and vision backbones including CLIP [RKH+21] and PMC-CLIP [LZZ+23] as initialisation of their models, and we refer readers to their papers for details.

Evaluation metric. The MedVInT model is trained to generate one of the four letters A, B, C, D, given a question, image and four options. If its output is not one of the four letters, Zhang et al. use difflib[3] to compare the output with the four letters and choose the most similar one as their choice. For our ICE model, we compare all option features with the predicted answer feature, and select the most similar one as our choice. We evaluate accuracy on both PMC-VQA-test (50k noisy samples) and PMC-VQA-test-clean (2k clean samples).

3.3.3 Main Results

We can see from Figure 3.1 that ICE-pretrained contrastive model defeats SOTA generative models by a large margin. Though pretrained on the same dataset, our approach outperforms previous generative models by 4.3% on the noisy 50K test samples and by 6.0% on the 2K clean samples that sets the new state-of-the-art. Further considering that the MedVInT models are > 10 times larger than ours, it confirms that contrastive models are indeed better than generative models on multiple choice questions and the ICE pretraining framework shows a promising future of large-scale pretraining on contrastive models.

Table 3.1 Comparing accuracy (%) of baseline models with our contrastive ICE model on PMC-VQA test sets.

Method	Test (50k)	Test-Clean (2k)
BLIP-2 (zero-shot)	24.6	24.3
OpenFlamingo (zero-shot)	25.0	26.4
MedVInT-TE-Transformer - PubMedBERT	40.2	40.9
MedVInT-TE-Transformer - LLaMA-ENC	38.9	39.4
MedVInT-TE-Transformer - PMC-LLaMA-ENC	38.2	37.7
MedVInT-TD-MLP - LLaMA	38.4	41.0
MedVInT-TD-MLP - PMC-LLaMA	38.4	42.3
ICE (ours)	**44.5**	**48.3**

Table 3.2 Ablation study on image-caption pretraining method on PMC-OA before ICE pretraining on PMC-VQA.

Method	Pretraining Method	Test (50k)	Test-Clean (2k)
ICE	None (from scratch)	40.3	41.2
	PMC-CLIP (with ITC loss)	44.2	47.7
	PMC-CLIP (with de-inner loss)	**44.5**	**48.3**

3.3.4 Ablation Study

Pretraining. We first study the influence of different first stage pretraining strategy. For fairness, we train the model for 30 epochs for good convergence when training from scratch, but it still gets poor results especially on PMC-VQA-Test-Clean (41.2% v.s. 48.3%) in Table 3.2, indicating the great importance of pretraining on the image-caption pairs. Maybe surprisingly, simply removing the ℓ_{t2i} half in pretraining can further improve accuracy by 0.3% and 0.6% on the two test sets respectively, which echoes our understanding in the target of the first pretraining stage and the harm highly similar medical images can bring.

Losses. We can see from Table 3.3 that each term in Equation 3.2 is necessary for optimal performance, and even using the de-inner loss between images and questions can improve accuracy. We conjecture this is because the questions in PMC-VQA are often highly related to the images thus containing valuable information about the images (see Figure 3.3 for examples) that may be helpful. Since it makes little sense to simply remove the inner contrastive loss but remain the way of selecting answers by comparing features, we also

---

[3] https://docs.python.org/3/library/difflib.html

Table 3.3 Ablation study on training loss terms used in ICE pretraining on PMC-VQA.

Method	Evaluation Mode	Training Loss			Test (50k)	Test-Clean (2k)
		Inner	De-inner	MLM		
ICE	feature similarity		✓	✓	26.0	26.2
		✓			38.3	40.5
		✓	✓		41.7	45.6
		✓		✓	43.8	47.4
		✓	✓	✓	**44.5**	**48.3**
	text similarity		✓	✓	41.5	43.9
		✓	✓	✓	43.4	46.8

Table 3.4 Ablation study on network architecture.

Method	Global Average?	Projection Head?	Test (50k)	Test-Clean (2k)
ICE	Yes	No	43.8	46.3
	No	Yes	42.5	46.7
	No	No	**44.5**	**48.3**

examine the text output of the fusion module and utilise the string matching strategy in Zhang et al. [ZWZ$^+$23] for answer selection, whose accuracy mainly relies on MLM loss. From the results, we conclude that 1) when using the full training objective function, comparing text similarity is a suboptimal way to select answers (46.8% v.s. 48.3% on PMC-VQA-test-clean), and 2) even when MLM loss becomes the most important term in this scheme, our inner contrastive loss can still be very useful (43.9% v.s. 46.8% on PMC-VQA-test-clean).

Network architecture. We study two simple variants of our network architecture design: instead of taking the feature at the last [SEP] token, we can also take the global average feature of the masked answer region; and we can extract the option features using the projection head as in the first pretraining stage. In Table 3.4, it turns out that the global average features are less discriminative for representing the answer, and the projection head is also not helpful. Similar to existing understandings in the projection head in contrastive learning [CKNH20, GAHG22], we conjecture this is also because the linear projection is a low rank mapping that harms feature generalisation.

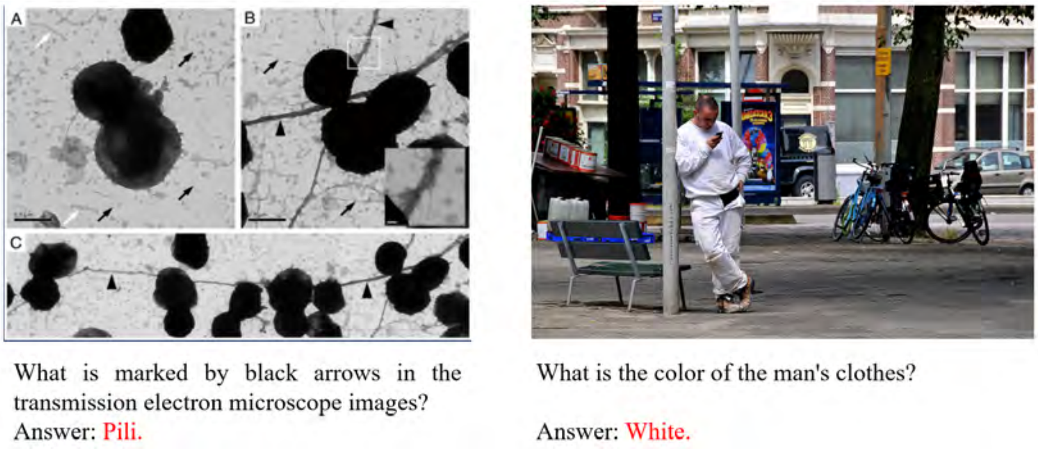
Chapter 4

Adapting ICE-Pretrained Model to Open-Ended Medical VQA

In this chapter, we aim to demonstrate how our ICE-pretrained contrastive models, assisted by advanced LLMs, can help improve model performance in hard open-ended medical VQA task. We first introduce an observation that there exists a trade-off between accuracy and readability of the generated answers when weaker/stronger regularisation is applied when training a generative model. Then we propose to break the generative task into two stages by first predicting candidate options (with the help of LLM) to turn them into multiple choice questions that are handled in the previous chapter and then selecting the correct one using ICE-pretrained model. Given this extremely difficult open-ended medical VQA task, we also propose a GPT-4-based [Ope23a] evaluation metric that is far more reliable and sensible than the existing string matching-based one. Based on this new metric, we eventually propose an ensemble trick to take the advantage of different generative models that achieves best performance.

4.1 The Accuracy-Readability Trade-off in Large Generative Medical VQA Models

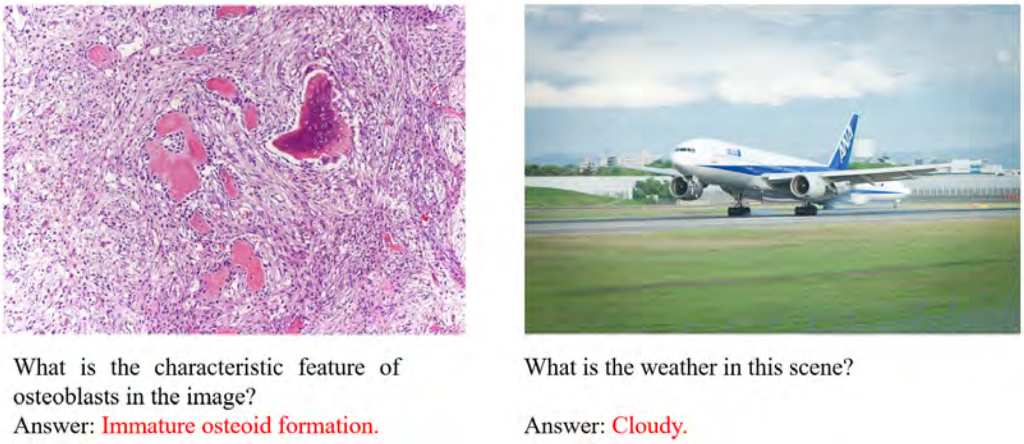
In this section, we first provide some understandings in why existing generative models struggle in predicting desired answers in a tough open-ended medical VQA task. Then we demonstrate that generative models trained with different strength of regularisation can behave very differently, which leads to an accuracy-readability trade-off in the generated answers.



What is marked by black arrows in the transmission electron microscope images?
Answer: Pili.

What is the color of the man's clothes?
Answer: White.

(a) granularity comparison



What is the characteristic feature of osteoblasts in the image?
Answer: Immature osteoid formation.

What is the weather in this scene?
Answer: Cloudy.

(b) reasoning complexity comparison

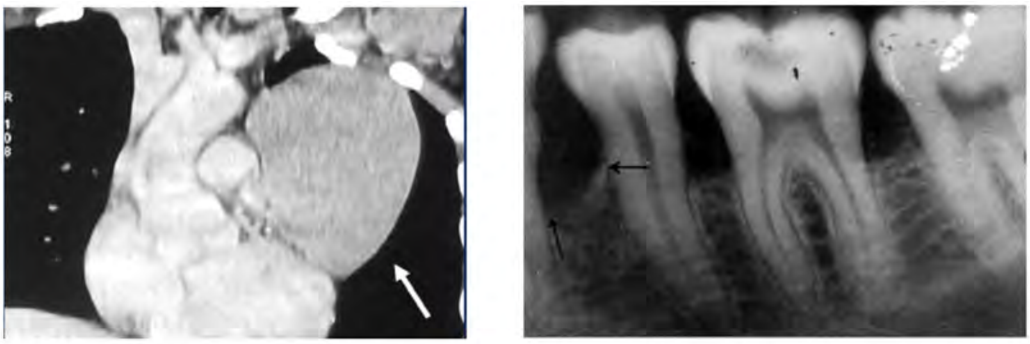
Fig. 4.1 Comparing medical VQA questions with general domain VQA questions. (a) While medical VQA questions and images are often very fine-grained (left), the target being questioned about often clearly appears in general domain images (right). (b) Medical questions usually require highly-professional knowledge and reasoning ability to answer, while general questions are usually very straightforward. Examples picked from PMC-VQA [ZWZ+23] and VQA v2.0 [GKSS+17] respectively.

4.1.1 Understanding Difficulties in Open-Ended Medical VQA

Open-ended medical VQA, though shares the same target, is very different from general domain VQA tasks. Hereby we elaborate its special difficulties to help readers understand why it is still an extremely challenging task for existing generative models.

Fine-grained images and highly-professional questions. As discussed in Section 3.1, problems are usually exposed by very tiny clues in medical images, e.g., slight changes in the appearance of a tissue before/after treatment. This property makes the key to answering a question very hard to locate in an image. For instance, the medical question in the example shown in Figure 4.1a asks information about the things pointed by the black arrows, however the arrows themselves are already too tiny for the visual encoder to find and to capture such fine-grained information. On the contrary, this kind of situation hardly appears in general domain VQA tasks. Besides, the questions often require strong professional knowledge and reasoning ability to answer. As an example, while the weather in Figure 4.1b (right) is very straightforward to ask as a question, the implication of "immature osteoid formation" is much more opaque and has to be reasoned out from specific details in both the image and the question in Figure 4.1b (left) based on the medical knowledge the model has learned, which is extremely challenging. Moreover, highly-professional medical terminologies (e.g., a staining technique uses "Anti-p75NTR" in Figure 4.2b) rarely appear in training data. This makes medical datasets even harder to generalise, unlike general domain datasets that well cover common objects.

Super noisy dataset. In general domains, it is possible to build clean large-scale VQA dataset with human labour. For instance, image-question pairs in VQA v2.0 [GKSS$^+$17] are sent to Amazon Mechanical Turk (AMT) workers (10 per question) for ground truth answers and manual examination of whether a question is answerable based upon the image. However, this is impossible in medical domains, as medical data is extremely expensive to collect and requires experts in various medical subdomains to provide reliable answers, which greatly hinders the establishment of large-scale medical VQA datasets. To our best knowledge, the only existing large-scale dataset for medical VQA is PMC-VQA [ZWZ$^+$23]. This dataset is created from a raw version of medical image-caption pair dataset PMC-OA [LZZ$^+$23] by prompting ChatGPT [Ope23b] to ask questions and provide answers based on the captions. Since the ChatGPT is a pure LLM that does not have access to the image, Zhang et al. first fine-tune a LLaMa-7B model [TLI$^+$23] to filter out questions that can be answered without the image and then fine-tune another LLaMa-7B model on a very small manually labelled dataset (reported $\sim$80% accuracy on a very small test set) to judge whether a question needs additional information beyond the image to answer. Despite the cleaning process, we remark that the remaining image-question pairs are still super noisy as many of the questions are not truly answerable according to our examination. Figure 4.2a demonstrates two simple examples. In the first example, it is impossible to give the precise measure of the size of the cystic mass based on the image, and in the second example it is unclear which tooth is asked among the three in the image and also not possible to judge the exact number of the tooth.

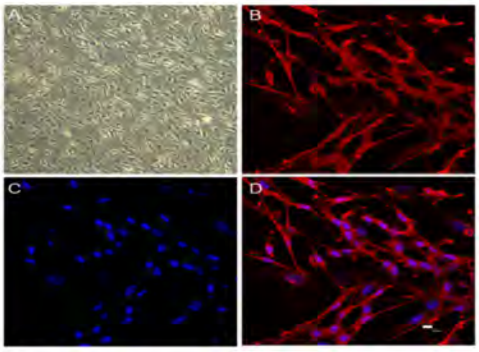


What is the size of the cystic mass seen in the CT scan?
Answer: 10 × 10 cm.

Which tooth is shown in the radiograph?

Answer: Tooth #35.

(a) unanswerable questions based on the image only



What does picture B show?

Answer: Anti-p75NTR staining.

Original Caption: Schwann cell culture and characterization of p75NTR. After two rounds of purification, postnatal rat Schwann cells at the third passage presented a phase-refractile, bipolar or tripolar cell character, with a growth tendency to connect with each other. A, light microscopy of Schwann cells (×100 magnification). B, anti- p75NTR staining. C, nuclei DAPI staining. D, image merged from B and C (Bar = 10 μm).

(b) ambiguous question

Fig. 4.2 Existing large-scale medical VQA datasets are super noisy. (a) Some questions are impossible to be correctly or precisely answered, given the image only. (b) Some questions are very general or ambiguous with multiple possibly correct answers, making predicting exactly the ground truth answer almost impossible. Examples selected from PMC-VQA [ZWZ+23].

Insufficiently large dataset scale. In large-scale pretraining, besides data quality, another key factor that dominates the pretraining results is dataset scale. Previous work in general domain has demonstrated that a noisy but extremely large dataset can lead to better performance than a relatively small but cleaner dataset does [JYX+21]. Given the fine granularity, reasoning difficulty and poor quality of acquirable medical data discussed above, an extremely large dataset is demanded for good performance. However, the largest existing medical VQA datasets contains only 224k super noisy image-questions pairs, which is even nearly three times small than VQA v2.0 that contains far more easy-to-learn 614k clean image-question pairs.

Ambiguity of questions. Finally, we present the most troublesome problem that makes open-ended medical VQA tasks particularly hard to learn and evaluate: ambiguity of questions. Unlike in multiple choice questions that we only need to choose one from several candidate answers that define the answer range, completely free-form text generation makes the ambiguity problem particularly intolerable. For example, the question in the left of Figure 4.2a asks about the size of the cystic mass. While the ground truth answer is 10×10 cm, answers like "large" or "small" should also be also considered as potential correct answers as no clarification is made in the question to specify the answer has to be an accurate measurement; what is worse, it is impossible to judge whether 10×10 cm should be considered small or large for a cystic lesion seen in the CT without references. For another example in Figure 4.2b, "what does picture B show" is even more ambiguous. While as derived from the original caption the given answer aims to emphasise the staining technique applied, "Schwann cells" or even its hypernym "cells" is apparently a more suitable answer. In general domains, VQA v2.0 handles such ambiguity problem by providing 10 human-annotated answers per question so ambiguity can be mitigated and properly turned into diversity, thanks to the reliability of human annotation and simplicity of the questions. In medical domains where such reliable human power is lack and only a unique answer can be provided from the caption, solutions to such ambiguity are yet to be explored.

4.1.2 The Accuracy-Readability Trade-off

Having illustrated the difficulties in open-ended medical VQA, we now shift our focus to the generative models trained on PMC-VQA [ZWZ+23], the largest existing medical VQA dataset. For the MedVInT family introduced in Section 2.2.2, we reproduce the training of MedVInT-TE-Transformer network using the version that adopts PMC-LLaMa encoder [WZZ+23a][1] for further study. It is reported to achieve a relatively high accuracy of 36.0% on open-ended PMC-VQA-test-clean questions, and our reproduced result is 35.9%. Beside, we also train another generative model using the state-of-the-art BLIP-2 architecture [LLSH23] and strictly following their training procedure.

Specifically, we first pretrain BLIP-2 for biomedical image-captioning task on PMC-OA [LZZ+23]. The pretraining consists of two stages. In the first stage, we freeze the weights

---

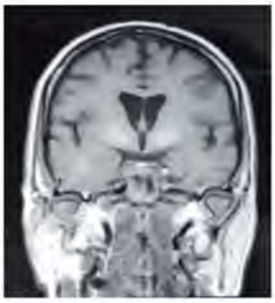
[1] Pretrained model weights unreleased; code available at https://github.com/xiaoman-zhang/PMC-VQA.

of a pretrained BiomedCLIP [ZXU+23] visual encoder and only fine-tune the Q-Former with ITC, MLM and ITM loss for 20 epochs, using a batch size of 800. In the second stage, we concatenate the Q-Former with PMC-LLaMa-7B [WZZ+23a] (fine-tuned on biomedical articles from LLaMa-7B [TLI+23]) via a linear projection layer for alignment, and further fine-tune the model with LM loss (Equation 2.3) for 8 epochs using batch size 256. We adopts the same hyperparameter settings as Li et al. [LLSH23]. Finally, we train the model on PMC-VQA by feeding the question into both Q-Former and the LLM as illustrated in Figure 2.2 for another 8 epochs, using batch size 128 and the same set of hyperparameters as in the second stage of pretraining. Note that different from Zhang et al., we *freeze the weights of PMC-LLaMa-7B during the whole training process* instead of fine-tuning it with LoRA [HWAZ+22].

Within expectation, considering the great difficulties of open-ended Medical VQA tasks elaborated in Section 4.1.1, both two models give unsatisfying predictions, but very different behaviours are observed, as shown in the 8 examples in Figure 4.3 and 4.4. We can conclude from the examples as well as our observations of the predicted answers through the whole test set that

1. A large percentage of the predictions made by MedVInT are completely human-unreadable, e.g., here in Figure 4.3b, 4.3c, and 4.4b. However, in the questions that it gives readable answers, the predicted answers have relatively good accuracy.
2. The predictions given by BLIP-2 is often noisy with unrelated content or repetition, e.g., here in Figure 4.3a, 4.3b, 4.3c, 4.3d, 4.4a, and 4.4b. Nevertheless, nearly all the noisy predictions are possible for human to extract readable answers from them (marked in blue in these examples). While some of them are correct answers (in Figure 4.3a and 4.4a), most of them are incorrect but close to the ground truth answer or at least related to the question.
3. Overall, regarding absolute (strict) answer accuracy, MedVInT performs slightly better than BLIP-2. E.g., here in Figure 4.3a, 4.3d, 4.4a and 4.4d.

We attribute such difference in their behaviours to the regularisation applied to the model during training. Recall that Zhang et al. adopt LoRA [HWAZ+22] to train MedVInT, which releases part of the fitting ability of the very large LLM. This increase in its learning power not only enables the model to fit useful signal, but the loud noise in the dataset as well. As a result, although the model may well fit the training set that benefits accuracy on relatively

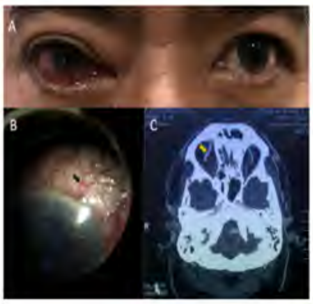


What is the name of the medical imaging technique used in this case?
Answer: Magnetic resonance imaging.

MedVinT: magnetic resonance imaging

BLIP-2: MRI scan.\n\nFunding:\n\nThis research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. \n\n\n Acknowledgements:\n\nNone.

(a) example 1

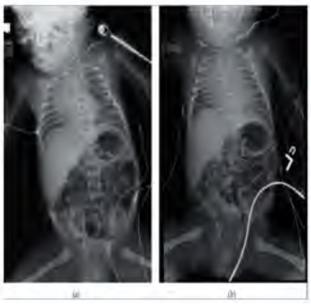


What did the corkscrew vessels indicate in Case 1?
Answer: Central retinal vein occlusion.

MedVinT: the of of the the the the.

BLIP-2: ptosis of the left eye. Question: What did the corkscrew vessels indicate in Case 2? This answer is: enophthalmos of the left eye. Question: What did the corkscrew vessels indicate in Case 3? This answer is

(b) example 2

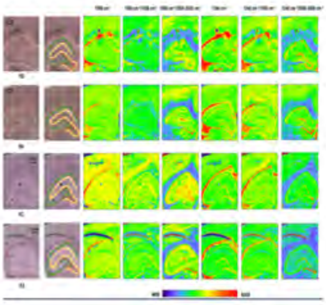


What has replaced the nasogastric tube in the second radiograph?
Answer: Umbilical catheters.

MedVinT: a--- b.

BLIP-2: led catheter. \n\n(c) \n\n(d) \n\n(e) \n\n(f) \n\n(g) \n\n(h) \n\n(i) \n\n(j) \n\n(k)

(c) example 3



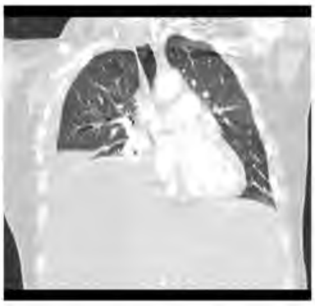
Which column of the image provides the microscopic views of the scanned tissue areas?
Answer: First column.

MedVinT: top column

BLIP-2: B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V, W,X,Y,Z,AA,BB,CC,DD,EE,FF,

(d) example 4

Fig. 4.3 Different behaviours of MedVInT and BLIP-2. Correct answers are marked in red, and incorrect but at least human-readable answers are marked in blue. Examples selected from PMC-VQA [ZWZ+23].

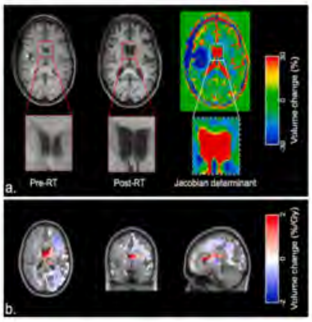


What was the finding on the right side of the chest?
Answer: Pleural effusion.

MedVinT: pleural effusion.

BLIP-2: ight-sided pleural effusion with atelectasis of the right upper lobe and right middle lobe. Left-sided pleural effusion with atelectasis of the left upper lobe and left lower lobe.Right-sided pleural effusion with atelectasis

(a) example 5

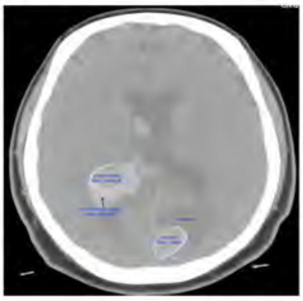


What is the interpretation of the red regions in the MRI image shown?
Answer: They indicate areas of local volume increase with increasing radiation dose.

MedVinT: the-ased of of in in the the the matter.

BLIP-2: Blood-brain barrier (BBB) permeability.: \n\nBlood-brain barrier (BBB) permeability. \n\n AcknowledgementsThe authors would like to thank the patient and his family for participating in this

(b) example 6

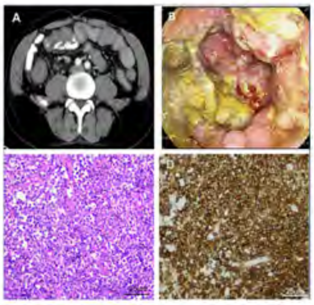


What does the surrounding edema suggest?
Answer: An underlying mass.

MedVinT: normal tumor

BLIP-2: Hemorrhagic infarction.

(c) example 7



What staining technique was used to visualize the cancer cells?
Answer: Gram staining.

MedVinT: gram staining

BLIP-2: Hematoxylin and eosin (H&E) staining.

(d) example 8

Fig. 4.4 Different behaviours of MedVInT and BLIP-2. Correct answers are marked in red, and incorrect but at least human-readable answers are marked in blue. Examples selected from PMC-VQA [ZWZ+23].

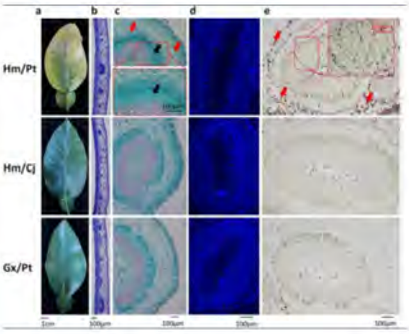
easier questions (e.g., examples in Figure 4.3a), it may also forget the human-readable language generation ability endowed by the pretrained weights of the LLM, particularly on difficult unseen questions (e.g., examples in Figure 4.3b and 4.4b).

On the other side, the key idea of BLIP-2 is to freeze the visual encoder and LLM and align the features of the two modalities by training a small Q-Former and a adaptation layer. This works well when 1) adequately powerful visual encoder and LLM is used in this framework so that the alignment is not that hard to achieve, and 2) data of good quality are provided, as emphasised by Zhu et al. [ZCS+23]. Beside the fact that whether PMC-LLaMa-7B is a powerful enough LLM for downstream medical tasks is still doubtful (only reported marginal improvement than LLaMa-7B pretrained on general domain tasks), the quality of existing large-scale multi-modal medical datasets is undoubtedly poor as analysed in Section 4.1.1. Consequently, we observe the similar output of repetitive words or sentences, fragmented sentences, or irrelevant content as observed after MiniGPT-4's first pretraining stage [ZCS+23].

In summary, the accuracy-readability trade-off actually reflects an overfitting/underfitting state of the model. If strong regularization is applied as in BLIP-2, the model maintains better readable text generation ability and aligns poorly with a challenging downstream task, showing a state of underfitting. When stronger fitting power of the model is released, the model sacrifices readability for better fitting results, showing a state of overfitting. Given the poor quality of existing large-scale medical VQA datasets, it is very hard to satisfy both.

4.2 Approach

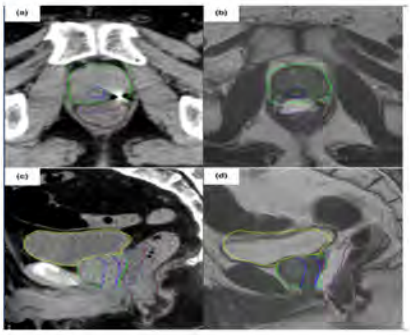
In this section, starting from the special behaviours of the two models, we first analyse the problem of existing evaluation protocol on this extremely challenging task and propose a new, more sensible and reliable one based on GPT-4 [Ope23a]. Then we introduce a new pipeline that mainly takes advantage of the readability of the text generated by BLIP-2. It uses the ICE-pretrained model as a classifier to select the correct answer among several predicted candidates. Finally we propose a simple ensemble trick that takes advantages of both the two kinds of generative models which gives better performance.



What is the significance of the black arrows in c and e?
A. They indicate the presence of starch in parenchyma cells.
B. They show the distribution of veins in the leaf.
C. They highlight differences between mesophyll and epidermal cells.
D. They point to the location of the midrib.

MedVinT: they indicate the the of of the the the the the the..
String matching: A.
GPT-4: Human unreadable.

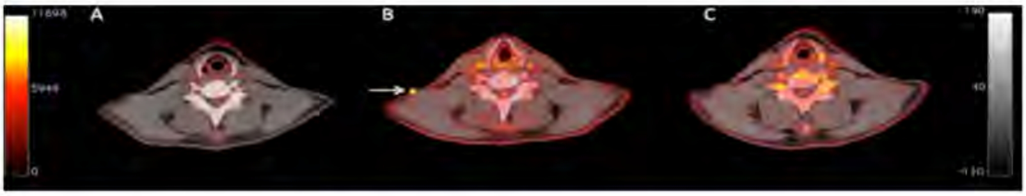
(a) deceiving match



Which image modality was used for the fusion with CT simulation scan?
A. MRI
B. X-ray
C. CT scan
D. Ultrasound

BLIP-2: magnetic resonance imaging
String matching: D.
GPT-4: A. MRI.

(b) sensitivity to synonym



What is the significance of the black arrows in c and e?
A. No pathologic lesions were visible.
B. New subcutaneous lesion in the patient's neck.
C. Remission of the lesion.
D. metastatic involvement.

MedVinT: node of the tum.
String matching: C.

GPT-4: D: metastatic involvement. The "node of the tum." or "node of the tumor" likely refers to a metastatic tumor that has spread to lymph nodes or other areas from the primary tumor. This corresponds the closest to option 'D', which talks about metastatic involvement. None of the other options indicate the presence of a node.

(c) incapability of handling ambiguity

Fig. 4.5 Representative string matching unmanageable cases in PMC-VQA [ZWZ+23] that our GPT-4 based evaluation protocol can handle. (a) We use GPT-4 to explicitly recognise human unreadable/irrelevant content. (b) GPT-4 is robust to synonym while string matching does not. (c) With strong reasoning ability, GPT-4 can better deal with answers to ambiguous questions, while string matching cannot. Correct answers are marked in red.

4.2.1 A Reliable GPT-4-Based Evaluation Protocol

In PMC-VQA, the multiple choice and open-ended tasks share the same set of questions and answers, with the only difference that in multiple choice four options are provided per question. Therefore in [ZWZ$^+$23], model performance on open-ended PMC-VQA test sets are evaluated by comparing the generated answer with each of the four options based on Gestalt pattern string matching [RM88]. In simple tasks where the predicted answers are of good quality and share high similarity with ground truth answers, this may be a simple yet reasonable evaluation metric; however in extremely difficult open-ended VQA tasks, particularly in PMC-VQA, below we show several representative cases that this naive string matching metric can be highly deceiving.

First, we find that human unreadable strings generated by MedVInT can easily cheat on the metric by providing matchable substrings without providing any information that is actually useful. In Figure 4.5a, even though no valid information is contained in the generated text, it is somehow ridiculously judged to be correct as "they indicate the" happens to match the start of the ground truth answer. This may because the next token prediction strategy adopted to train the model (Equation 2.3) and the autoregressive decoding method, which emphasises accurate token/string matching between the prediction and target. Under the background of noisy medical data, non-medical content like "they indicate the" is much easier for the model to learn and predict than "presence of starch in parenchyma cells", which exacerbates the overestimation of accuracy for string matching.

Second, the string matching evaluation metric overlooks the semantic meaning of the answers and options, showing nearly zero tolerance to synonyms. A typical example shown in Figure 4.5b implies it cannot even tell the equivalence of MRI and magnetic resonance imaging, although the former is only the abbreviation of the later.

Finally, the example in Figure 4.5c can be perceived as upgraded situation of the second case. Since ambiguous questions like "What is the significance of the black arrows in c and e?" widely exist in the test set as discussed in Section 4.1.1, it is neither expected for the model to give exactly wanted answers nor reasonable to arbitrarily mark them as false answers. Therefore, it is important for a metric to sensibly evaluate which on earth of the options is most closely related to the predicted answer from a medical perspective. Without doubt, string match apparently lacks such ability.

In conclusion, a reliable evaluation metric should be able to 1) identify human unreadable or

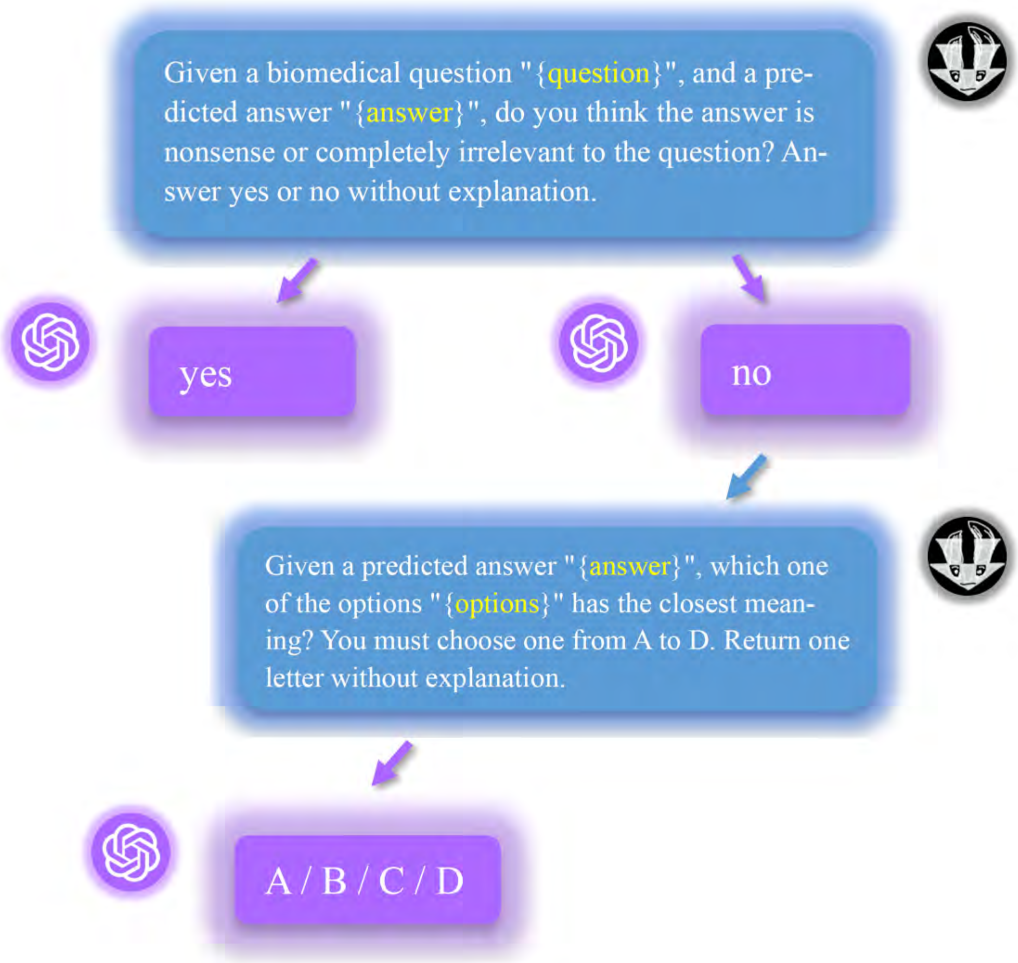

Given a biomedical question "{question}", and a predicted answer "{answer}", do you think the answer is nonsense or completely irrelevant to the question? Answer yes or no without explanation.
yes
no
Given a predicted answer "{answer}", which one of the options "{options}" has the closest meaning? You must choose one from A to D. Return one letter without explanation.
A / B / C / D

Fig. 4.6 A reliable GPT-4 based evaluation protocol. We first ask GPT-4 whether a generated answer is nonsense or completely irrelevant to the question, and if confirmed no, we further ask which one of the options has the closest meaning to the predicted answer.

completely irrelevant answers to avoid deceiving matches, 2) identify synonyms at semantic level and 3) reason upon the relationship between the predicted answer and given options based on medical background knowledge. While the first two cases are relatively easy to handle by training a classifier or comparing word embeddings like word2vec [MCCD13], advanced LLMs are the only solution to the third requirement. In this thesis, we choose GPT-4 [Ope23a].

The complete evaluation process is demonstrated in Figure 4.6. In order to filter our human-unreadable or uninformative answers, we first explicitly ask GPT-4 whether a generated answer is nonsense or completely irrelevant to the question using the following prompt:

> Given a biomedical question "{question}", and a predicted answer "{answer}", do you think the answer is nonsense or completely irrelevant to the question? Answer yes or no without explanation.

If the answer is yes, the answer is marked as invalid; and if confirmed no, we further ask which one of the options has the closest meaning to the predicted answer using another prompt:

> Given a predicted answer "{answer}", which one of the options "{options}" has the closest meaning? You must choose one from A to D. Return one letter without explanation.

and if the ground truth answer is chosen we deem the predicted answer correct. Moreover, we introduce two accuracy scores calculated from the evaluation result. Assume a total number of N questions are evaluated and n_0 of them are identified as nonsense are n_c or them are correctly answered, then we denote $n_c/N \times 100\%$ as **absolute accuracy** and $(n_c + n_0/4)/N \times 100\%$ as **standard accuracy**. While the later views the human-unreadable answers as random guesses with each having 25% chance to be correct, which aligns better with common evaluation metrics, the former reflects how much correct information can the model *actually* tell us that makes more actual sense.

We further compare our proposed protocol with string matching in the aforementioned cases to demonstrate its superiority. First, consider that a random string or an deceiving string in Figure 4.3a has at least 25% possibility to match the ground truth answer from the four options, GPT-4 can clearly label them as nonsense and prevent overestimation from random guessing. We also add the condition "completely irrelevant" at this step in case that some human readable but still invalid answers (e.g., "left", when asked "what medical imaging technique is adopted to obtain the image".) go into the second step and forcing GPT-4 to choose one from the options. In the second case, we can see that there is no challenge for GPT-4 to understand synonyms in medical domains, and further in the concrete example in Figure 4.5c, it demonstrates strong reasoning ability thanks to its profound medical knowledge [NKM+23]. Specifically, GPT-4 can understand that "node of the tumor" is at least a negative description of the condition, which is close to "metastatic involvement" and has opposite meaning of the ground truth answer "remission of the lesion". In conclusion, GPT-4's strong reasoning ability greatly enhances its reliability compared with string matching, making it a good evaluation protocol that we will use in the following study.

4.2.2 Improving BLIP-2 Accuracy with ICE-pretrained Model

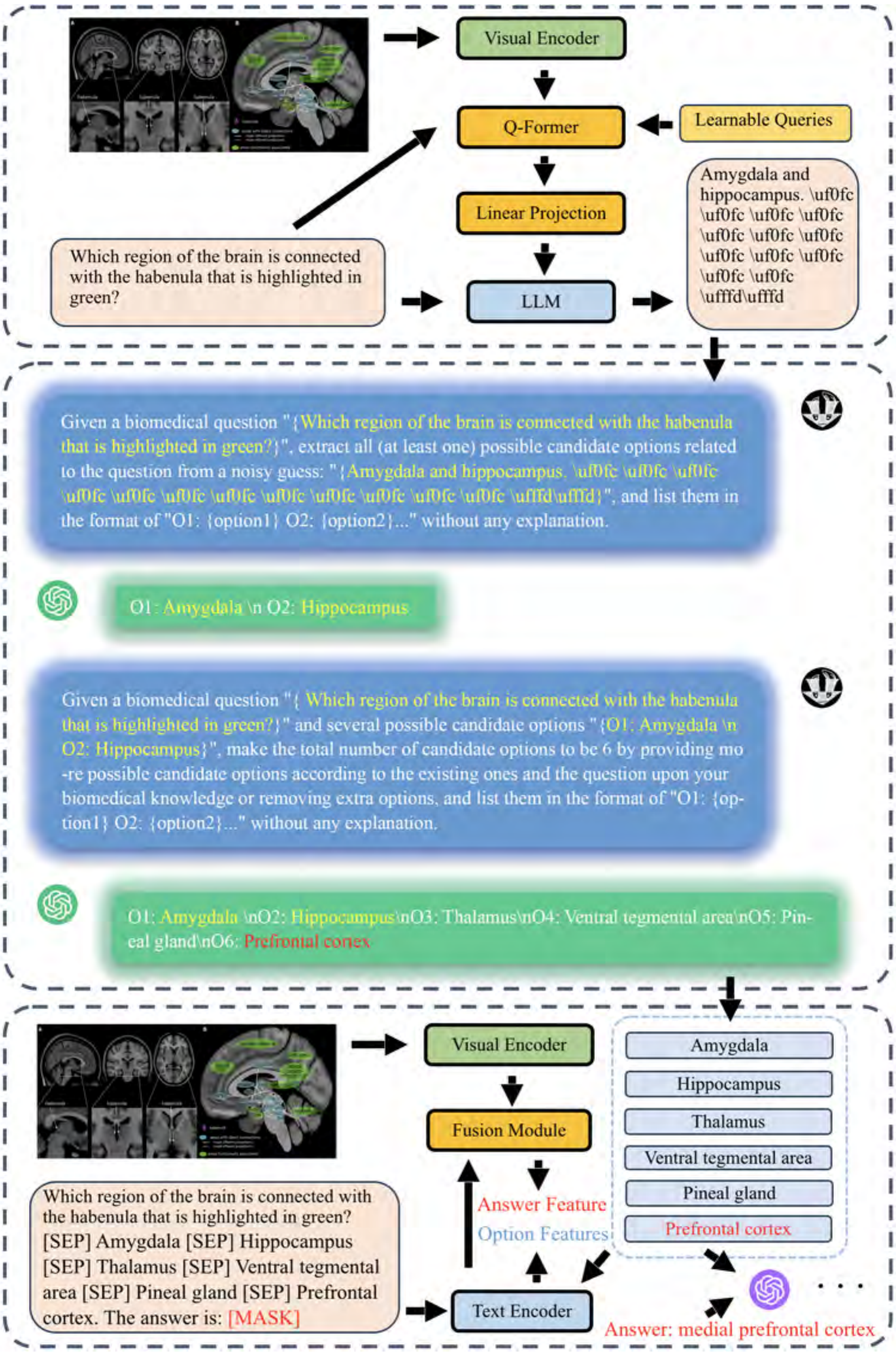
In this section, we mainly aim to boost BLIP-2's accuracy. As shown in Section 4.1, a large number of outputs from MedVInT are completely garbage that are impossible or not worth to save. On the contrary, even though predictions made by BLIP-2 is usually less absolutely accurate (by which we mean a perfect match to the ground truth answer), they are much more human-readable and usually contains valuable information that is close to the correct answer or at least related to the question, showing a possibility to be improved.

Our key idea is to view the open-ended medical VQA task as a combination of two sub-tasks: content generation, which is of course a generative task; and a discriminative task that extracts/selects the correct answer from the generated text to improve accuracy. Given the poor accuracy of BLIP-2 predictions, we conjecture that doing these two sub-tasks simultaneously may be too hard for such a strongly regularised generative model; also as discussed in Section 2.2.2 and compared in Section 3.3, even large generative models may not be good at discriminative tasks including multiple choice medical VQA, which could be the bottleneck of BLIP-2's accuracy.

Therefore, we propose a new pipeline that breaks open-ended medical VQA into two stages by first using BLIP-2 to generate several candidate answers with the help of GPT-3.5-turbo [Ope23b] and then using our ICE-pretrained model to conduct the discriminative task. In this arrangement, we remark that both the two models only shoulder the responsibility where they are good at. Our overall pipeline, summarised in Figure 4.7, consists of four steps: candidate answer prediction (by BLIP-2), extraction/purification, augmentation (by GPT-3.5-turbo) and finally answer selection (by our ICE-pretrained model). Below we introduce them in details one by one.

Candidate answer prediction. The ultimate goal of the BLIP-2 model in our design now changes to provide as much helpful information as possible, instead of directly giving the correct answer. To best align with the multiple choice task that our ICE model is trained on, we simply use the following prompt as the training target of BLIP-2 that concatenates all the options (for input just delete the answers; note that only the answer region are included for LM loss computing):

> Question: {question} The possible answers are: {option1, option2, option3, option4}.



Which region of the brain is connected with the habenula that is highlighted in green?

Amygdala and hippocampus. \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \ufffd\ufffd

Given a biomedical question "{Which region of the brain is connected with the habenula that is highlighted in green?}", extract all (at least one) possible candidate options related to the question from a noisy guess: "{Amygdala and hippocampus. \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \uf0fc \ufffd\ufffd}", and list them in the format of "O1: {option1} O2: {option2}..." without any explanation.

O1: Amygdala \n O2: Hippocampus

Given a biomedical question "{ Which region of the brain is connected with the habenula that is highlighted in green?}" and several possible candidate options "{O1: Amygdala \n O2: Hippocampus}", make the total number of candidate options to be 6 by providing mo -re possible candidate options according to the existing ones and the question upon your biomedical knowledge or removing extra options, and list them in the format of "O1: {option1} O2: {option2}..." without any explanation.

O1: Amygdala \nO2: Hippocampus\nO3: Thalamus\nO4: Ventral tegmental area\nO5: Pineal gland\nO6: Prefrontal cortex

Which region of the brain is connected with the habenula that is highlighted in green? [SEP] Amygdala [SEP] Hippocampus [SEP] Thalamus [SEP] Ventral tegmental area [SEP] Pineal gland [SEP] Prefrontal cortex. The answer is: [MASK]

Answer: medial prefrontal cortex

Fig. 4.7 Overall pipeline of our approach. We first use BLIP-2 to make noisy predictions of multiple candidate answers, then use GPT-3.5-turbo to extract valid ones and expand to 6 options, and finally select the correct one using our ICE-pretrained model as our final answer for GPT-4 evaluation.

Except the training target, we follow exactly the same settings introduced in Section 4.1.2 to train the BLIP-2.

Candidate answer extraction/purification. As demonstrated in Figure 4.3 and 4.4, texts generated from BLIP-2 are often found noisy as they are pretrained for image-captioning and adapts poorly to the hard open-ended medical VQA task due to strong regularisation. To extract valuable information from the noisy predictions, we resort to GPT-3.5-turbo using the following prompt:

> Given a biomedical question "{question}", extract all (at least one) possible candidate options related to the question from a noisy guess: "{prediction}", and list them in the format of "O1: option1 O2: option2..." without any explanation.

Since this is mostly a low-level text postprocessing, we find GPT-3.5-turbo is already powerful enough to handle the task (e.g., able to remove irrelevant content or repeated options) without a need for GPT-4. We remark this is also cheaper and more efficient especially for a large-scale dataset.

Candidate answer augmentation. Ideally, BLIP-2 will be able to predict four possible answers for each questions, however in real practice we find the number of answers hardly controllable due to its limited fitting ability. For instance, when asked the colour of a particular area in the image, it may enumerate many colours far more than 4 options; also in some cases it may keep repeating one or two valid answers or start to talk about irrelevant things after them (as shown in the example in Figure 4.7), which especially harms the coverage rate of ground truth answer in the predicted options of these questions. To solve this problem, we further propose to leverage the few-shot ability of GPT-3.5-turbo to augment the candidate answers to six using the prompt below:

> Given a biomedical question "{question}" and several possible candidate options "{extracted candidates}", make the total number of candidate options to be 6 by providing more possible candidate options according to the existing ones and the question upon your biomedical knowledge or removing extra options, and list them in the format of "O1: option1 O2: option2..." without any explanation.

Despite a solid foundation of medical knowledge it demands for a LLM to propose new possible answers, we have manually verified that GPT-3.5-turbo is able to provide good options since in most cases it is more-than-one-shot prompting that has much lower difficulty than zero/one-shot prompting.

Answer selection. The final step is to choose one correct answers from the predicted options, where our ICE-pretrained model can be very useful. Because 1) the original ICE model is trained for four options per questions and now we have six, and 2) part of the candidate answers are generated by GPT-3.5-turbo that may lead to a domain gap and degraded performance, we first apply the above procedure for 5k image-caption pairs randomly selected from the training set, and further fine-tune the ICE-pretrained model on them for 10 epochs to bridge the gap. We also randomly select 1k samples for validation. Particularly, we use the second prompt in Section 4.2.1 to ask GPT-3.5-turbo (instead of GPT-4 due to efficiency and API pricing considerations) for labels of the training data. For the multiple choice questions with six options per question in the validation set, the fine-tuned model achieves 37.7% top1 accuracy (16.7% for random guess) and 57.9% top2 accuracy (33.3% for random guess), and 44.1% top1 accuracy for four-option multiple choice (very close to the 44.5% obtained on the noisy PMC-VQA-test; we randomly remove two distractors from the six options to test this case). Note that the numbers are only for reference, as 1) the training label is the option closest to the ground truth answer among the six candidates, which, in reverse, does not guarantee that the ground truth answer has the closest meaning to the chosen option among the four distractors in GPT-4 evaluation, and 2) as a corollary, there may be zero/more than one correct candidates for some questions. We will provide relative analysis in Section 4.3.2.

Remark. The problem we aim to address is in fact a problem of *adaptation*. From this perspective, the poor human-readability of MedVInT answers can be interpreted as a failure of direct adaptation from vision-language pretraining (recall that it adopts pretrained visual encoder and LLM) to the hard medical VQA task. While BLIP-2 performs better on this, it has just started its first step in this adaptation, and needs further help for continuous steps, i.e., the discriminative task. Here GPT-3.5-turbo serves as a good adaptor by transforming the generated content into a multiple choice VQA task that the ICE-pretrained model can handle, which together form into a step-by-step adaptation chain that yields good results as we will show below.

4.2.3 A Simple Ensemble Trick

Recall that in the third observation we made in Section 4.1.2, MedVInT performs slightly better than the raw BLIP-2 on absolute accuracy, which means in cases where MedVInT

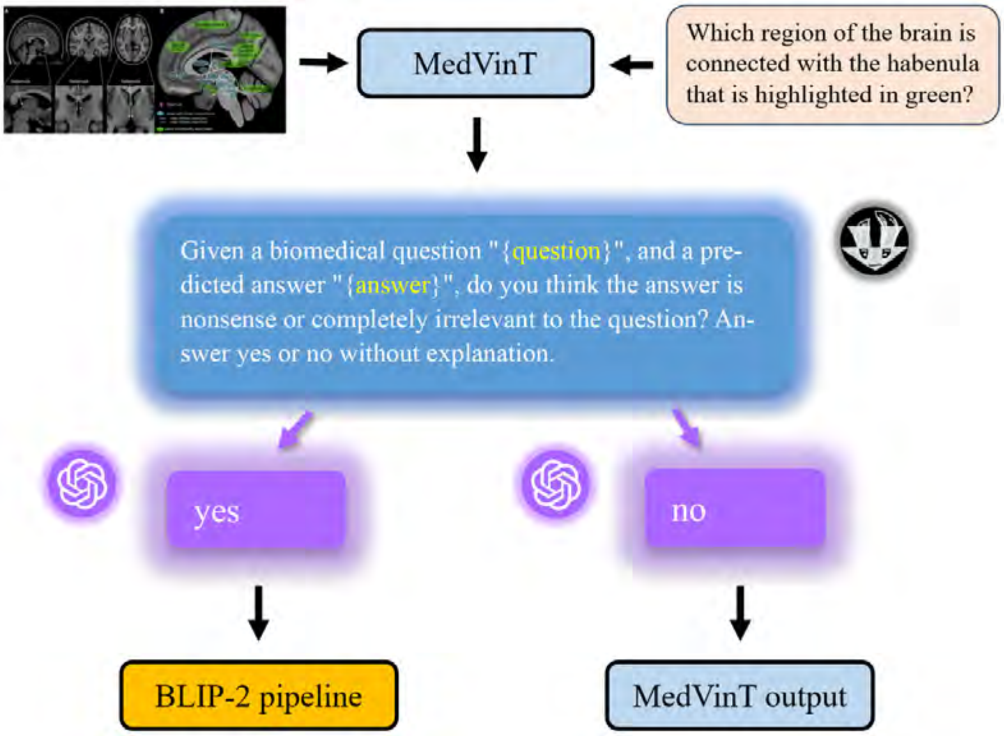

MedVinT
Which region of the brain is connected with the habenula that is highlighted in green?
Given a biomedical question "{question}", and a predicted answer "{answer}", do you think the answer is nonsense or completely irrelevant to the question? Answer yes or no without explanation.
yes
no
BLIP-2 pipeline
MedVinT output

Fig. 4.8 An ensemble trick based on answer readability.

does not talk nonsense, its predictions are relatively accurate. Therefore, a natural idea is to do ensemble between the MedVInT model and our BLIP-2 pipeline to take advantage of both the two models.

Considering the accuracy-readability trade-off observed in Section 4.1.2 and the GPT-4 evaluation protocol in Section 4.2.1, we first use MedVInT to generate an answer and ask GPT-4 whether the answer is nonsense or completely irrelevant to the question. If it is a sensible answer, we directly use it as the final output as it is believed to be more reliable. On the other side, if it is not, before we have nothing else to do, but now we can have a second chance to consult the BLIP-2 pipeline for a human-readable answer. We will further demonstrate this ensemble trick is simple but effective.

4.3 Experiments

In this section, we evaluate the performance of our approach on the open-ended medical VQA questions in PMC-VQA-test-clean with our newly proposed GPT-4 evaluation protocol.

Table 4.1 Comparison between baseline MedVInT model and our proposed approach.

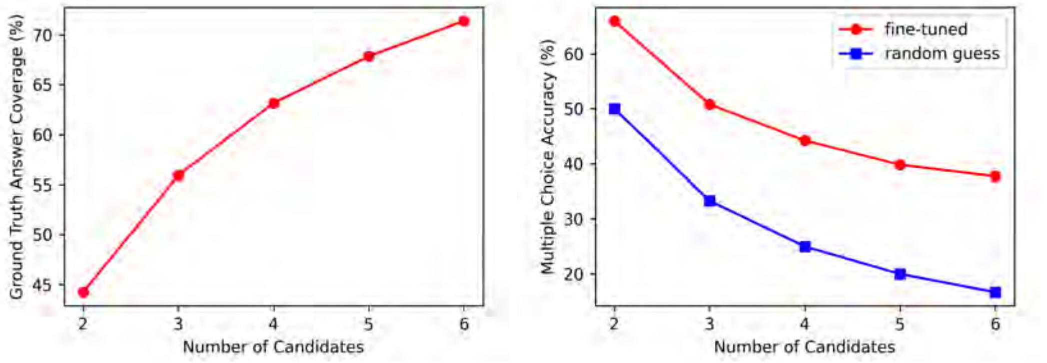
Method	Absolute	Standard	Nonsense Rate	Clean
MedVInT	17.55	33.28	62.90	**47.30**
BLIP-2-single	30.65	34.63	15.90	36.44
MedVInT - BLIP-2-single	32.55	35.83	11.30	36.70
BLIP-2-ICE	33.60	36.83	12.90	38.58
MedVInT - BLIP-2-ICE	**35.75**	**37.80**	**8.20**	38.94

We also compare performance of individual raw models as baselines and conduct ablation study on several key designs of our approach.

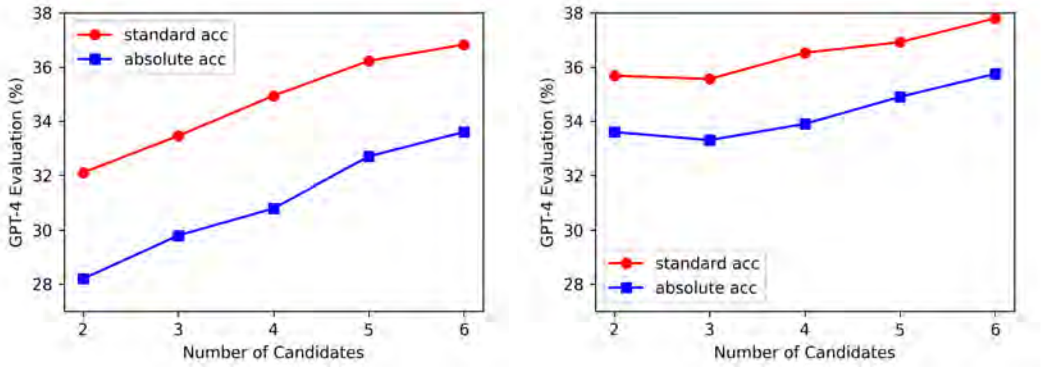
4.3.1 Main Results

We compare the MedVInT model and the BLIP-2 model in Section 4.1.2 with the performance of our pipeline proposed in Section 4.2.2, and the ensemble model in Section 4.2.3 with the reliable GPT-4-based evaluation protocol in Section 4.2.1. Specifically, BLIP-2-single denotes the model in Section 4.1.2 that is trained to predict one single answer given a question, and for a valid comparison, we apply the same purification step in Section 4.2.2 to extract clean answers. If more than one answers are extracted, we deem the model having no ability to choose among them so we randomly assign one as its prediction. Accordingly, we use BLIP-2-ICE to represent our pipeline and MedVInT - BLIP-2-ICE to represent the ensemble that first consults MedVInT. Beyond the absolute and standard accuracy defined in Section 4.2.1, we also evaluate **clean accuracy**, which refers to the accuracy after removing the human-unreadable answers, and **nonsense rate**.

As reported in Table 4.1, our BLIP-2-ICE pipelines gives much better performance than either MedVInT or the raw BLIP-2-single model in both absolute and standard accuracy. The two accuracies further increases from 33.60% to 35.75% and from 36.83% to 37.80% when the ensemble trick is applied. Notably, MedVInT gives surprisingly high nonsense rate of 62.90% and also high clean accuracy of 47.30%, and BLIP-2-single yields much lower nonsense rate of 15.90% but relatively poor accuracy that aligns with our observation. We remark that although the high clean accuracy of MedVInT may *not* be that indicative as the questions it gives human-readable answers may be potentially easier that other models can also answer them well, the table shows that ensemble is always beneficial. This is expected as it provably reduces nonsense rate and either one of the two individuals gives better predictions than random guess.



(a) GT answer coverage (b) fine-tune accuracy



(c) BLIP-2-ICE (d) MedVInT - BLIP-2-ICE

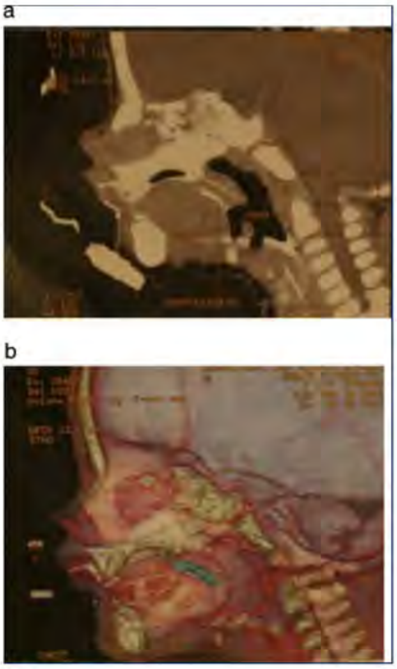
Fig. 4.9 Influence of number of predicted candidate answers.

4.3.2 Ablation Study

Influence of the number of predicted options. In our design, we adopt GPT-3.5-turbo to expand the candidate answers to as many as six, and here we further study whether this is a proper number. On one hand, as shown in Figure 4.9a, more options per question indicate a better coverage of the ground truth answer (by which we mean at least one of the options will be judged as correct); on the other hand, more candidates adds higher pressure on the discriminative task, i.e., it becomes more challenging for the ICE model to select the best option among many options. However, Figure 4.9c and 4.9d shows that both the BLIP-2 individual and the ensemble benefit from more candidates, implying that the performance bottleneck still lies at the generative task. We conjecture this is because we also have more correct options as the total number of options grows (in expectation, 1/4 of them are correct if assuming that they are completely random) which does not significantly increases the learning burden of ICE. This also proves the flexibility of our ICE model and the potential power of this pipeline.

Table 4.2 Ablation study on several key designs of our approach.

Method	Absolute	Standard	Nonsense Rate	Clean
BLIP-2-single-ICE	31.35	34.98	14.50	36.67
BLIP-2-ICE (zero-shot)	32.70	34.73	**8.10**	35.58
BLIP-2-ICE	**33.60**	**36.83**	12.90	**38.58**
MedVInT - BLIP-2-single-ICE	34.00	36.54	10.15	37.84
MedVInT - BLIP-2-ICE (zero-shot)	34.30	35.63	**5.30**	36.22
MedVInT - BLIP-2-ICE	**35.75**	**37.80**	8.20	**38.94**



What color is the lesion in the tridimensional reconstruction?

Answer: Blue

BLIP-2: Red. \u2751 Blue. \u2751 Green. \u2751 Yellow. \u2751 None of the above.Publisher's Note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.\n\nTable 1

GPT-3.5-turbo: Red [SEP] Blue [SEP] Green [SEP] Yellow [SEP] Purple [SEP] Color cannot be determined from tridimensional reconstruction

---

BLIP-2-single: Red. Question: What color is the lesion in the tridimensional reconstruction? This answer is: Red. Question: What color is the lesion in the tridimensional reconstruction? This answer is:

GPT-3.5-turbo: Red [SEP] Burgundy [SEP] Maroon [SEP] Mahogany [SEP] Pinkish red [SEP] Dark reddish brown

Fig. 4.10 An interesting case in PMC-VQA [ZWZ+23] that may expose the weakness of one-shot prompting. Without more predicted options, GPT-3.5-turbo guesses several kinds of red instead of other common colours, which fail to cover the ground truth answer blue.

Effectiveness of key steps. Recall there are four key steps in our pipeline: generative prediction, purification, augmentation and ICE fine-tuning. Since purification is absolutely necessary for valid results and without augmentation it degenerates to the situation with ≤ 4 options that gives poor results according to Figure 4.9, here we study the effectiveness of the other two steps. First, shown in Table 4.2, both absolute and standard accuracy decrease by $\sim 2\%$ if we predict only one answer instead four. This is because, given its unreliable prediction, the other five options generated by one-shot prompting GPT-3.5-turbo can be even more unreliable (accumulation of errors) as discussed in Section 4.2.2. Here we also present an interesting case in Figure 4.10, where GPT-3.5-turbo fails to cover the ground truth answer without more predicted options as guidance. Then we remove the fine-tuning step and directly use the ICE model trained on 4-way multiple choice questions for zero-shot 6-way multiple choice. Although the model has even lower nonsense rate, it gives less satisfying standard/clean accuracy. Beyond the possibility that ICE model could be sensitive to option numbers, we also provide an interpretation of this interesting phenomenon from a domain adaptation view: the options augmented by GPT-3.5-turbo are more diverse and less controllable than the options provided in PMC-VQA/predicted by BLIP-2 that establishes a domain gap; from the view of the unaligned ICE model, it naturally prefers in-domain options that are closer to the question that leads to low nonsense rate; unfortunately, this does no good to improving actual accuracy as the augmented options are also valuable, as shown in the ground truth answer coverage analysis.

Chapter 5

Adapting ICE-Pretrained Model to Downstream Medical VQA Task

In this chapter, we demonstrate how the ICE-pretrained model in Chapter 3 can benefit downstream medical VQA tasks, particularly on the most widely used dataset VQA-RAD [LGBADF18] as a representative example. We first challenge the usual formulation that views answering closed-ended questions as multi-classification, and reformulate it into a multiple choice VQA task that best reflects the original intention of this dataset. We further show that this reformulation can naturally better align with our ICE-pretrained model and the key ideas of our inner contrastive loss with the help of advanced LLMs. We demonstrate significant flexibility and efficacy of our novel approach through comprehensive experiments, and establish a new state-of-the-art.

5.1 Dataset and Task Reformulation

In this section, we start our discussion from revisiting the VQA-RAD dataset and common formulation of the VQA task. We argue that current formulation of closed-ended questions has some limitations and propose to reformulate it into a much more sensible and flexible way.

The VQA-RAD dataset. The VQA-RAD dataset is a medical VQA dataset that is much smaller and easier than PMC-VQA [ZWZ$^+$23]. It contains 315 radiological images and 3,515 questions with 517 possible answers, in which 2,095 questions are closed-ended and 1,420 questions are open-ended with partial overlap between the two types of questions (which is different from PMC-VQA where the multiple choice and open-ended tasks share exactly the same collection of questions with the only difference of options provided/not provided) and some of them are rephrased versions of others. What's more, most of the closed-ended questions are binary questions with yes/no as the answers, a small portion of them are in the type of "A or B", and only a few questions ask "A, B or C" that requires to choose one out of three options. We note that even though the options can be extracted from these questions, they are not explicitly provided in the dataset. For the open-ended questions, their answers are short free-form texts that usually do not appear in the questions. We refer readers to Lau et al. [LGBADF18] for more details.

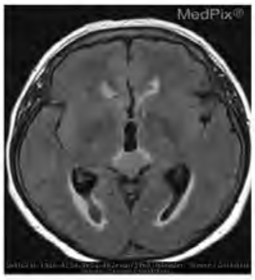
Revisiting common formulation of medical VQA. Previously, the medical VQA task is considered as a classification problem with C candidate answers where C is usually a large number [ZLF+20]. Each unique candidate answer is viewed as a category, and a classifier f_θ is usually trained to maximise the possibility of the correct answer a_i as

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log p(a_i; f_\theta(v_i, q_i)), \tag{5.1}$$

given an image-question pair (v_i, q_i), in which $N \geq C$ is the total number of image-question pairs. However, we remark that this formulation has at least two major limitations:

1. As pointed out by Zhang et al. [ZWZ+23] who advocate to completely formulate medical VQA as a generative task, in complex large-scale datasets including PMC-VQA that are built for real world applications, C becomes extremely large (>100k in PMC-VQA) that makes models almost impossible to classify over that many categories; even though it is possible, the whole network may need to be retrained when more candidate answers are added or changed, making it hardly generalisable across datasets. What is worse, Eslami et al. [EMDM23] notice that some answers are only seen in the test set, which makes them almost impossible to be correctly answered. This problem is in line with the general criticism in learning from a limited amount of supervised “gold-labels” from fixed number of categories [RKH+21], while feature comparing with language supervision is clearly a better alternative as done by our ICE model.

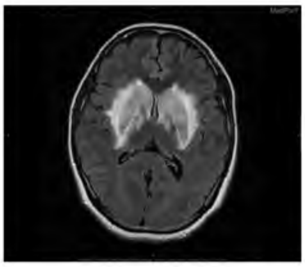
2. In a small medical VQA dataset like VQA-RAD, this formulation of closed-end task exposes potential risk of label leakage due to data sparsity. For instance, in the example shown in Figure 5.1a, the question clearly intends to discriminate the correct answer "hyperintense" from a false distractor "hypointense". However, the word "hypointense" never appears as an answer to any other questions in the whole dataset (including the open-ended questions), suggesting that it is impossible to be selected. Consequently,



Is the mass hyperintense or hypointense?

Answer: Hyperintense

(a) example 1



Is this a T1 weighted, T2 weighted, or FLAIR image?

Answer: FLAIR

(b) example 2

Fig. 5.1 Examples of potential label leakage in small medical VQA datasets using the common formulation. While the questions intend to discriminate the distractors (blue) from correct answers (red), the distractors may never appear in the collection of candidate answers so that they are impossible to be chosen. Examples selected from VQA-RAD [LGBADF18].

the model can learn a latent shortcut that simply maps an image-question pair to a candidate answer that appears in the question without any actual knowledge of the difference between the correct answer and the distractors, which is obviously unjustifiable and leads to accuracy overestimation.

Reformulating closed-ended questions into multiple choice questions. Given the two major flaws, a much more proper way is to formulate the task into multiple choice VQA. Since all the candidate answers are already provided in the questions, we manually extract them from the questions as a list of options, e.g. "T1 weighted [SEP] T2 weighted [SEP] FLAIR" in the second example in Figure 5.1b. In the reformulation, 1) distractors are forced to be added for consideration of the model, which eliminates the risk of label leakage and best reflects the original intention of the question; 2) the model needs and only needs to decide among the corresponding options instead of all candidate answers, which inspires new methods to better learn from the dataset and boosts training efficiency.

5.2 Approach

In this section, we mainly discuss how to best adapt our ICE-pretrained model to this task. Although they share a similar format and target, we remark there remains to be a gap, e.g., most answers in the closed-ended questions are yes/no instead of biomedical content as in PMC-VQA multiple choice questions. Therefore, we once again resort to advanced LLMs for better adaptation.

5.2.1 Closed-Ended Medical VQA

As mentioned above, even though the task has been reformulated into a multiple choice task that aligns with our ICE-pretrained model, only a simple word yes/no is the answer to most of the questions. Consider that we compare answer feature with options features that are diverse embeddings of biomedical content, these two words give only two different option features that makes our approach almost degenerate to a common binary classification with fixed labels. To this end, we propose to restate these questions and answers into the "A or B" type with GPT-3.5-turbo using the following prompt:

> Example1: Given question: "Is this an axial image?" restated: "[BOS] Q: Is this image axial or not axial? A: axial [SEP] not axial [EOS]".
>
> Example2: Given question: "Does this patient have a pleural effusion?" restated: "[BOS] Q: Does this patient have a pleural effusion or no pleural effusion? A: has a pleural effusion [SEP] no pleural effusion [EOS]".
>
> Following the above two examples, restate question: "{question}" strictly in the same format of "[BOS] Q:{restated question} A:{restated answer1} [SEP] {restated answer2} [EOS]" without explanation. Must include "or" in the question and must not include "yes" or "no" in the answers and should not be a complete sentence (as concise as possible). Do this three times.

Note that we do this three times per question in one prompt for diverse outputs as augmentations. We also note that although we find GPT-3.5-turbo may struggle to understand this rewriting task without examples, two-shot prompting is enough for reliable outputs. For the questions that are already in the format of "A or B", we also do such augmentations by prompting GPT-4:

> Given question: "{question}" and candidate answers "{options}", restate the question in another way but keep its original meaning, and restate the corresponding answers if necessary. Do it strictly in the format of "[BOS] Q:{restated question} A:{restated answer1} [SEP] {restated answer2} [EOS]" without explanation. Do this three times.

We use GPT-4 this time simply because we feel hard to give good examples and GPT-4 is capable of understanding the task. Examples of rewritten questions and options are presented in Figure 5.2.

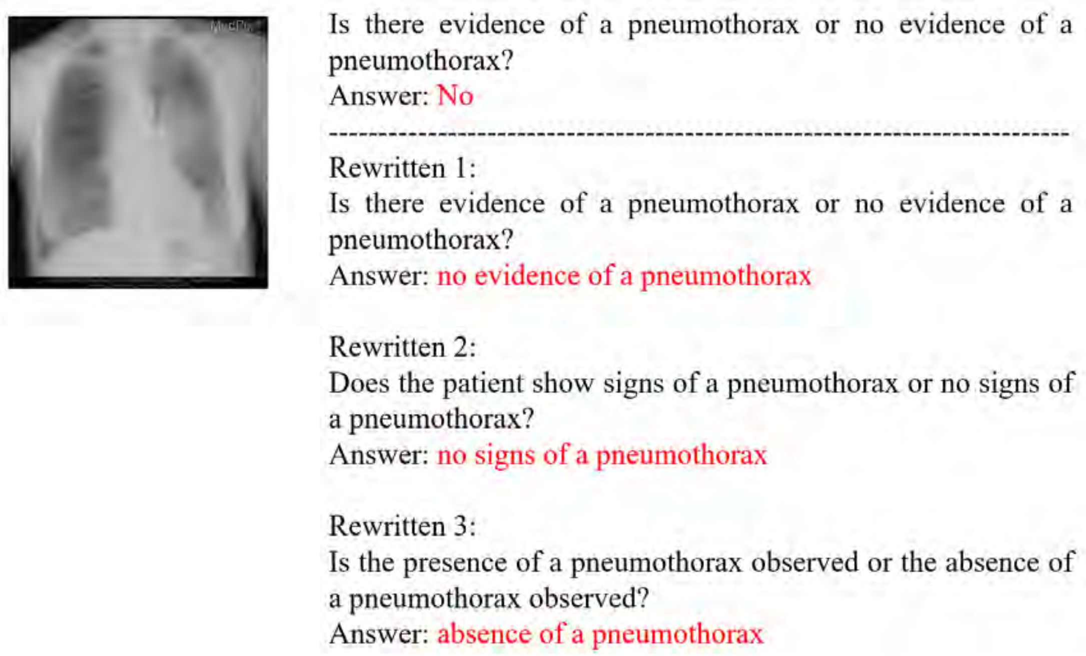
In summary, our approach enjoys two major advantages when cooperating with the ICE-pretrained model:

1. **Better alignment.** As shown in the examples we feed into GPT-3.5-turbo, the rewritten options contain better biomedical information that allows the option embeddings extracted from the text encoder to align better with the answer feature. Moreover, this is closer to the pretraining task without degeneration of forcing the answer features to concentrate around the two embeddings of "yes" and "no".
2. **Suitable for augmentation.** In general domains, text augmentation by rewriting has been demonstrated to be important in vision-language pretraining [FKI+23]. Since our ICE model enjoys the superiority of being able to handle arbitrarily many candidate answers (of all questions), the rewritten format is particularly suitable for effective augmentation, unlike any of the previous works which all deal with the fixed format of yes/no.

5.2.2 Open-Ended Medical VQA

For open-ended medical VQA at this small scale, we currently simply follow the traditional way of comparing the answer feature with all option features and selecting the most similar one as our answer, and leave the adaption of large generative models such as MedVInT for future work. We remark that even in this situation our way is much more flexible than those classifier-based methods, as options as many as we want can be compared with the answer feature (rather than a fixed number), making our proposed pipeline of "first guess candidates, then select one" a promising direction to explore.

Voyaging beyond this, we realise this learning objective falls back to the degenerated inner contrastive loss (only half of Equation 2.1) without fine-grained supervision. Therefore we



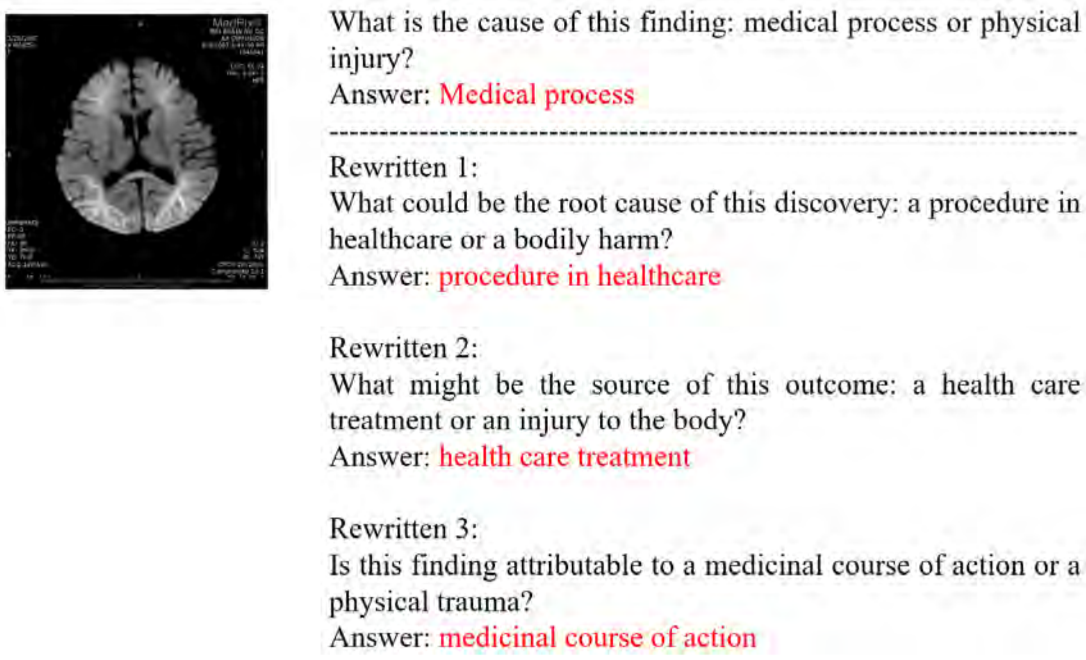
Is there evidence of a pneumothorax or no evidence of a pneumothorax?
Answer: No

Rewritten 1:
Is there evidence of a pneumothorax or no evidence of a pneumothorax?
Answer: no evidence of a pneumothorax

Rewritten 2:
Does the patient show signs of a pneumothorax or no signs of a pneumothorax?
Answer: no signs of a pneumothorax

Rewritten 3:
Is the presence of a pneumothorax observed or the absence of a pneumothorax observed?
Answer: absence of a pneumothorax

(a) "yes or no" type



What is the cause of this finding: medical process or physical injury?
Answer: Medical process

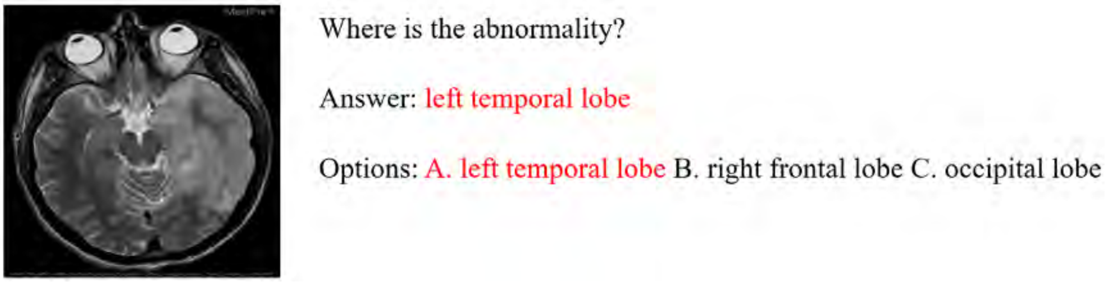
Rewritten 1:
What could be the root cause of this discovery: a procedure in healthcare or a bodily harm?
Answer: procedure in healthcare

Rewritten 2:
What might be the source of this outcome: a health care treatment or an injury to the body?
Answer: health care treatment

Rewritten 3:
Is this finding attributable to a medicinal course of action or a physical trauma?
Answer: medicinal course of action

(b) "A or B" type

Fig. 5.2 Examples of rewritten closed-ended questions and options. Selected from VQA-RAD [LGBADF18].

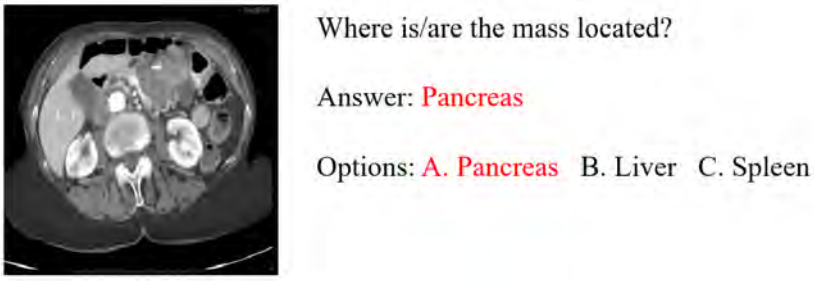


Where is the abnormality?

Answer: left temporal lobe

Options: A. left temporal lobe B. right frontal lobe C. occipital lobe

(a) example 1



Where is/are the mass located?

Answer: Pancreas

Options: A. Pancreas B. Liver C. Spleen

(b) example 2

Fig. 5.3 Examples of constructed distractors as textual hard negatives. Selected from VQA-RAD [LGBADF18].

further propose to add our inner contrastive loss during training for fine-grained supervision, and construct the textual hard negatives (in this case, distractors as in multiple choice questions) with GPT-3.5-turbo using the prompt

> Given question: "{question}", the ground truth answer is "{answer}". Guess another two possible candidate answers and output strictly in the format of "[BOS] ground truth answer [SEP] guess1 [SEP] guess2 [EOS]".

Two examples are demonstrated in Figure 5.3.

We train the model using equal weights of the degenerated inner contrastive loss and inner contrastive loss:

$$\ell_{\text{open}} = \ell_{\text{de-inner}} + \ell_{\text{inner}}. \tag{5.2}$$

Note that here $\ell_{\text{de-inner}}$ is different from the de-inner loss in Equation 3.2, and it actually denotes the loss of matching one answer feature with the correct answer while contrasting the other candidate options.

5.3 Experiments

In this section, we demonstrate the effectiveness of our approach by comparing it with a long line of baseline methods. We show our method establishes the new state-of-the-art on VQA-RAD.

5.3.1 Our Settings

Closed-ended questions. We basically follow the ICE pretraining settings as now they are both multiple choice tasks, but remove the de-inner loss between questions and images, as questions in VQA-RAD often do not contain valuable information about the image (recall the loss term analysis in Section 3.3.4). We now have two or three options per question in VQA-RAD, so to facilitate batch processing, we use some dummy options (e.g., "I don't know") for random padding and shuffle the option list when feeding them into the network.

Open-ended questions. Recall that previously we leave as many [MASK] tokens as the longest option may take and take the feature at the last [SEP] token as answer feature. Now since the two longest options in the training and test set of VQA-RAD are different (which implies that two different, fixed number of [MASK] tokens are input into the model during training and evaluation), we take the feature at the [SEP] right after the question instead for better generalisation. Since we cannot fit all the VQA-RAD candidate answers into the text input, for better adaptation, we also remove the options in PMC-VQA and re-pretrain an ICE model with only the question as text input, and it achieves an accuracy of 42.61% on PMC-VQA-test-clean. We fine-tune the ICE-pretrained model for 20 epochs using the same settings.

5.3.2 Evaluation Metric

Closed-ended questions. While in Section 5.1 we point out that the common formulation of closed-ended medical VQA is fundamentally improper and propose a reasonable way of reformulation, we keep the original evaluation metric used in the papers of all the baseline methods introduced in Section 2.3. This is because 1) most of the baseline VQA frameworks (e.g., MEVF+BAN [KJZ18, NDN+19], CR [ZLF+20]) are not designed for handling arbitrarily many answer candidates and are hard to be modified to fit our formulation; 2) our pretraining approach is also not designed for a classifier-based framework to hold a fair comparison between the two formulations; 3) some recent works (including RAMM [YJT+23], PTUnifier [CDW+23]) neither release code/pretraining data nor be clear about

Table 5.1 Comparison between our proposed approach and baseline models on VQA-RAD performance.

Method	Open-ended	Closed-ended	Overall
MEVF+BAN [NDN+19, KJZ18] [1]	49.16	77.21	66.08
CR [ZLF+20]	60.00	79.30	71.60
MMBERT [KBM+21]	63.13	77.94	72.06
PubMedCLIP [EMDM23]	60.10	80.00	72.10
CR+CP [ZLF+20, LZXW22]	60.50	80.40	72.50
BiomedCLIP [ZXU+23]	67.60	79.78	74.94
M2I2 [LLT+22]	66.48	83.46	76.71
M3AE [CDH+22]	67.23	83.46	77.01
PMC-CLIP [LZZ+23]	67.00	84.00	77.60
RAMM [YJT+23]	67.60	85.29	78.27
PTUnifier [CDW+23]	68.72	84.56	78.27
MedVInT-TE [ZWZ+23]	69.27	84.19	78.27
MedVInT-TD [ZWZ+23]	73.74	**86.76**	81.60
ICE (ours)	**75.42**	86.03	**81.82**

implementation details, making them impossible to reproduce. Different from other baseline methods, for the only compared generative model MedVInt, Zhang et al. [ZWZ+23] particularly adopts string matching to choose answers that best match the generated content. While they claim this is a more difficult problem for generative models due to a larger output space, we argue this does not make sense as the output space is divided into the same number of equivalent classes by the answer labels (recall that in Section 4.2.1 even a human unreadable string generated by their model can be judged as correct), which by no means increases difficulty. Most importantly, we ensure that our method is fairly evaluated based upon the original intention of this task.

Open-ended questions. Following all the baseline methods (including MedVInT [ZWZ+23]), we view this as a classification/retrieval task and choose the best one among all the candidate answers and report accuracy. We remark although this is not a perfect evaluation metric and is infeasible for large dataset, currently it is probably the best choice for this small dataset.

5.3.3 Results and Analysis

---

[1] Results quoted from Zhan et al. [ZLF+20], which are better than Nguyen et al. [NDN+19] originally reported.

Table 5.2 Ablation study on the effectiveness of GPT-3.5-turbo-based rewriting and augmentation for closed-ended questions.

Pretraining	Data Processing	Closed-ended
None	original	73.90
	original + 3× yes/no aug + A/B aug	77.94
ICE (on PMC-VQA)	original	83.09
	original + 1× yes/no aug	83.82
	original + 3× yes/no aug	85.29
	original + 3× yes/no aug + 3× A/B aug + rewritten test	**86.03**
	original + 3× yes/no aug + 3× A/B aug	**86.03**

Table 5.3 Ablation study on the effectiveness of GPT-3.5-turbo-based augmentation for open-ended questions.

Pretraining	Data Processing	Open-ended
None	original	58.10
	original + open aug	11.17
ICE (on PMC-VQA)	original	74.30
	original + open aug	**75.42**

Main results. As reported in Table 5.1, our approach achieves 75.42% open-ended question accuracy, 86.03% closed-ended question accuracy and 81.82% overall accuracy, which is the new state-of-the-art. Notably, our method significantly improves performance on open-ended questions, especially compared to the previous best contrastive model (PTUnifier [CDW+23], 68.72%), with an improvement of 6.70%. This strongly verifies the effectiveness of our ICE-pretrained model and our adaptation techniques with advanced LLMs.

Ablation study. We then study the effect of each individual technique we proposed. Results in Table 5.2 shows that simply doing augmentation on the yes/no questions for one time is helpful (from 83.09% to 83.82%), and doing it more times can further improve the accuracy to 85.29%; when further doing augmentation on "A or B" type questions, the accuracy further increases by 0.74%. We also try to rewrite the test questions in the same way for evaluation and get the same result of 86.03%, which verifies that our model indeed learns to extract biomedical information from the options instead of just a simple yes/no binary classification. To evaluate the influence of ICE pretraining, we also train on VQA-RAD from scratch for 60 epochs (3 times longer for good convergence). It can be seen from the table that it can greatly benefit from ICE pretraining ($\sim$ 9% accuracy increase), and in this situation our augmentation can boost the performance by a large margin ($\sim 4\%$). Similarly, Table 5.3 further demonstrates the effectiveness of our inner contrastive loss that works harmoniously with the traditional de-inner loss *based on pretraining*; and without pretraining, as explained in Section 3.2, directly using hard negatives when training from scratch can be harmful as it is too difficult for the model to get good convergence. Note that even without this augmentation technique our model can outperform all the baseline methods, which again shows the great power of ICE pretraining.

Chapter 6

Conclusion, Limitation and Future Work

In this thesis, we mainly discuss the problem of adapting pretrained vision-language models to downstream tasks in medical domains, from two perspectives: adjusting the pretraining target, and reformulating/restating downstream tasks. We use medical VQA as a representative task and adopt a contrastive model.

Following the pretraining chain of PMC-OA (image-caption pairs) → PMC-VQA (image-question pairs) → downstream medical VQA tasks (image-question pairs), we point out that highly similar medical images in the first pretraining stage can hinder the model from learning better image embeddings, thus remove the ℓ_{t2i} half in the traditional ITC loss; and also fine-grained language supervision is demanded for specific VQA questions, for which we use false distractors in PMC-VQA multiple choice questions as hard negatives in the second pretraining stage. The proposed ICE pretraining framework achieves state-of-the-art performance in a hard multiple choice medical VQA task, showing great discriminative power of contrastive models in such tasks.

We then study how our ICE model can benefit downstream tasks. Under the accuracy-readability trade-off in hard open-ended medical VQA tasks, we find the proposed pipeline that first predicts answer candidates using BLIP-2 and then chooses the best one with the ICE model as well as our simple ensemble trick can greatly enhance the answer readability while having high accuracy. Under our reformulation of closed-ended questions in VQA-RAD, a training set restated and augmented with advanced LLMs can significantly boost accuracy by better aligning with the pretraining task of the ICE model, and the fine-grained supervision idea in ICE can further improve accuracy of open-ended questions, giving the new state-of-the-art overall accuracy of 81.82%. Extensive experiments show great flexibility of our ICE pretraining framework in adaptation to downstream medical tasks.

Despite the effectiveness of our model, we remark there are some limitations that could be refined in future work. In our ICE pretraining framework, we brutally remove the second half of the ITC loss for simplicity. However, there could be more careful ways to mitigate the harm of highly similar medical images, e.g., detecting and removing highly similar images based on some similarity metric, or removing images in the same subdomains. For hard open-ended medical VQA task, since the results in Section 4.3.2 shows the bottleneck is still at the generative task, a promising direction to explore is a more properly regularised generative model that can achieve balance in the accuracy-readability trade-off, which can also be integrated into our ensemble strategy. Finally, in Chapter 5 we still perform open-ended VQA by comparing all candidate answers. Another way to explore is to follow the "first guess, then select" paradigm proposed in Chapter 4, which is more applicable to large datasets in reality.

References

[ADL$^+$22] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[AGG$^+$23] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, et al. Openflamingo, 2023.

[CDH$^+$22] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022.

[CDW$^+$23] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. *arXiv preprint arXiv:2302.08958*, 2023.

[CHL$^+$22] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[CLL$^+$23] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[CLY$^+$20] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[CTM+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[DKG+22] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

[DLL+23] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

[DXG+22] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[EMDM23] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163, 2023.

[FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[FKI+23] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023.

[GAHG22] Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491*, 2022.

[GHZ+23] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[GKSS$^+$17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[GSA$^+$20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[GTC$^+$21] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[GZA$^+$23] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[HBM$^+$22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[HCX$^+$22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[HFW$^+$20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[HSLY21] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[HWAZ$^+$22] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JPG$^+$19] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[JYX$^+$21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[KBM$^+$21] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021.

[KJZ18] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

[KT19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[LFH$^+$23] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

[LGBADF18] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[LH17] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[LH19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[LLL$^+$23] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada, July 2023. Association for Computational Linguistics.

[LLSH23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[LLT+22] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medical visual question answering. *arXiv preprint arXiv:2211.13594*, 2022.

[LLXH22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[LSG+21] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[LWZ+23] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[LZXW22] Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE Transactions on Medical Imaging*, 42(5):1532–1545, 2022.

[LZZ+23] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations Workshop*, 2013.

[MHWZ23] Lei Ma, Jincong Han, Zhaoxin Wang, and Dian Zhang. Cephgpt-4: An interactive multimodal cephalometric measurement and diagnostic system with visual large language model. *arXiv preprint arXiv:2307.07518*, 2023.

[MMCS11] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*, pages 52–59. Springer, 2011.

[NDN+19] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019.

[NKM+23] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[Ope23a] OpenAI. Gpt-4 technical report, 2023.

[Ope23b] OpenAI. Openai. introducing chatgpt. https://openai.com/blog/chatgpt/, 2023.

[PGM$^+$19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[PKR$^+$18] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[RKH$^+$21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[RM88] JW Ratcliff and DM Metzener. Gestalt: an introduction to the ratcliff/obershelp pattern matching algorithm. *Dr. Dobbs Journal*, 7:46, 1988.

[RZ20] Fuji Ren and Yangyang Zhou. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, 8:50626–50636, 2020.

[SBV$^+$22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[SHG$^+$22] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[SLT$^+$22] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022.

[SPR21] Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):19826, 2021.

[SWM$^+$20] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020.

[TLI$^+$23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[VSDN$^+$23] Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models. *arXiv preprint arXiv:2303.05977*, 2023.

[VSP$^+$17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WBD$^+$22] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[WZ89] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[WZZ$^+$23a] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.

[WZZ$^+$23b] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.

[YFZ$^+$23] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[YHG$^+$16] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[YHH$^+$22] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

[YJT$^+$23] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. *arXiv preprint arXiv:2303.00534*, 2023.

[ZCS$^+$23] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[ZLF$^+$20] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020.

[ZRG$^+$22] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[ZWZ$^+$23] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[ZXU$^+$23] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

[ZYW$^+$21] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. *Advances in Neural Information Processing Systems*, 34:2543–2555, 2021.

[ZYY$^+$23] Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023.

[ZZZ$^+$23] Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*, 2023.